Universidade do Minho

Escola de Engenharia

Departamento de Informática

# A Gestão da Qualidade dos Dados em Ambientes de *Data Warehousing* na Prossecução da Excelência da Informação

Alexandre Manuel Pereira Mendes da Costa

**Dissertação de Mestrado**

2006

# A Gestão da Qualidade dos Dados em Ambientes de *Data Warehousing* na Prossecução da Excelência da Informação

Alexandre Manuel Pereira Mendes da Costa

Dissertação apresentada à Universidade do Minho para obtenção do grau de Mestre em Sistemas de Dados e Processamento Analítico, elaborada sob orientação do Professor Doutor Orlando Manuel de Oliveira Belo.

2006

*Aos meus filhos e ao meu pai.*

# Agradecimentos

Ao Professor Orlando Belo, pela perspicácia e pragmatismo demonstrado na discussão do tema alvo de investigação e pela orientação saudável ao longo da dissertação, assente numa relação de confiança, respeito, partilha e sabedoria.

À Universidade do Minho, em especial, ao corpo docente do Curso de Mestrado em Sistemas de Dados e Processamento Analítico, pela transmissão de saberes que abriram as portas para a realização da presente dissertação.

À organização, não divulgada por razões de confidencialidade, que proporcionou a realização do projecto METRICWARE e que se consubstanciou no estudo de caso da dissertação.

Ao Francisco Sales e ao Alberto Mendes, pelo tempo dispendido e pelas observações pertinentes sobre o trabalho realizado.

À Maria Madalena, pelo apoio total e compreensão manifestada desde o primeiro momento. Em duas palavras: por tudo.

À Virgínia Pereira por me ter feito despertar o interesse na frequência do curso de mestrado e pelo suporte logístico e material durante alguns períodos mais críticos do processo de realização deste trabalho.

Aos *The Mission*, por serem fieis companheiros de viagem durante a realização desta dissertação, sobretudo, pelos momentos de descontracção e recomposição de energias que me proporcionaram.

Por fim, a todos aqueles que, apesar de não serem aqui citados, contribuíram directa ou indirectamente para a realização deste trabalho.

# Resumo


## A Gestão da Qualidade dos Dados em Ambientes de *Data Warehousing* na Prossecução da Excelência da Informação

Nos nossos dias, os Sistemas de *Data Warehousing* são um dos mais importantes instrumentos no panorama organizacional. A capacidade em gerar mais e melhores informações e indicadores aos agentes de decisão, flexibilizando e reduzindo o consumo de tempo e recursos no processo de interrogações ao repositório de dados, são algumas das características que distinguem estes sistemas e os catapultam para um patamar de destaque no suporte ao exercício de tomada de decisão. Os resultados apresentados pelo sistema são, naturalmente, influenciados pelos dados captados e residentes no *Data Warehouse*. A existência de índices de qualidade dos dados inferiores aos aceitáveis, ao longo das diversas camadas da arquitectura dos Sistemas de *Data Warehousing*, faz reconhecer o princípio *garbage in*, *garbage out*. Assim, interessa identificar as causas para a presença da fraca qualidade dos dados no sistema, de modo a estabelecer, por um lado, os meios de resolução das irregularidades verificadas durante a estada dos dados nos Sistemas de *Data Warehousing* e por outro lado, no sentido da prevenção dessas ocorrências. A assumpção da informação divulgada como produto-informação, detentor de características próprias e elaborado a partir de um sistema de processos específicos que transformam as matérias-primas, em vista a satisfação das necessidades e desejos dos consumidores finais, mostra ser uma iniciativa importante para a obtenção e disponibilidade de dados de elevada qualidade.

A presente dissertação pretende expor a problemática da qualidade dos dados em Sistemas de *Data Warehousing*, apresentando, de raiz, um conjunto de conceitos e terminologias básicas, bem como relevar as principais técnicas, metodologias, modelos e estratégias, passíveis de, concertadamente, se consubstanciarem ao que designamos por uma plataforma de um sistema de gestão

da qualidade dos dados em Sistemas de *Data Warehousing*. Esta plataforma procura prever a manutenção da qualidade dos dados ao longo das diversas camadas constituintes dos Sistemas de *Data Warehousing*. Adicionalmente, definiu-se um lote de métricas capaz de permitir a recolha de índices sobre a qualidade dos dados nas suas dimensões mais representativas. Estes índices, estrategicamente dispostos pelas diversas camadas dos Sistemas de *Data Warehousing*, possibilitam aferir sobre o grau de sucesso destes sistemas relativamente aos dados disponibilizados. Complementarmente, é apresentado um estudo de caso, realizado sobre o *Data Mart* duma organização real, em vista o reconhecimento da problemática da qualidade dos dados em Sistemas de *Data Warehousing*. O estudo de caso procura, baseado no trabalho realizado, fornecer um conjunto de iniciativas de recomendação para solucionar os problemas detectados e consequentemente promover o princípio da prevenção de erros e a melhoria contínua dos dados organizacionais.

# Abstract

## Data Quality Management in Data Warehousing Systems – Pursuing the Information Excellence

Nowadays, Data Warehousing Systems are one of the most powerful tools that we can find in organization. The ability in generating more and better information to decision making agents, improving user analysis skills and reducing querying efforts, are some of the distinguish characteristics of these systems that push them to the vanguard of decision support systems. All the results provided by a data warehousing system are influenced by data gathered in selective information sources and stored in their data warehouses. The existence of data quality indices lower than the acceptable ones throughout the data warehousing system architecture layers emerges the principle of “garbage in, garbage out”. So, it is not difficult to see that we need to identify properly, as soon as possible, all the potential causes that justify the presence of bad quality data in system’s data repositories - bad data generates bad business decisions. This will allows us to establish possible ways to attenuate the effect of such irregularities and define some strategies to avoid the causes for bad data. The assumption that the information provided can be seen as information product seems to be an important initiative for the attainment and availability of data with high quality levels, once it detains the adequate properties and is generated by specific oriented processes that transform the raw materials according the needs of data consumers.

This thesis approach the data quality issue in Data Warehousing Systems, presenting its basic concepts and terminology, as well as revealing the main techniques, methodologies, models and strategies in the field. As a direct result of this work we got, what we call, a consolidated conceptual platform for data quality management in a data warehousing system, intending to anticipate data quality maintenance throughout its architectural layers. Additionally, we also defined a set of met-

rics capable of capturing indices on data quality in its more representative dimensions. These metrics, strategically disposed over the layers of data warehousing systems, make possible to measure the degree of success of the data that they use to publish and provide to decision-making agents. Furthermore, a case study was carried out, based on a real world data mart, in order to identify and recognize the main data quality issues that we need to deal with when we face a real data warehousing system. The case study provided a very interesting set of recommendations that will allow us in the solution of data quality problems and, consequently, to promote the principle of error prevention and continuous improvement of organizational data.

# Índice

# Índice de Figuras

# Índice de Tabelas

# Siglas e Acrónimos

| | |
|---|---|
| ARD | Área de Retenção de Dados |
| CCDWS | Competence Center Data Warehousing Strategy |
| CDC | Change Data Capture |
| CIA | Central Intelligence Agency |
| CRM | Customer Resource Management |
| DM | Data Mart |
| DW | Data Warehouse |
| DWQ | Data Warehouse Quality |
| EAI | Enterprise Application Integration |
| ERP | Enterprise Resource Planning |
| ESB | Enterprise Service Bus |
| ETL | Extracção e Transformação e Carregamento |
| FBI | Federal Bureau of Investigation |
| GQM | Goal Question Metric |
| INS | Immigration and Naturalization Service |
| IPMAP | Information Product Map |
| MMLC | Measurement Model Life-Cycle |
| NASA | National Aeronautics and Space Administration |
| OLAP | On-Line Analytic Processing |
| OMB | Office of Management and Budget |

| | |
|---|---|
| PDCA | Plan Do Check Act |
| PI | Produto-informação |
| PSP/IQ | Product and Service Performance for Information Quality |
| QFD | Quality Function Deployment |
| ROI | Return On Investment |
| SDW | Sistema de Data Warehousing |
| SGBD | Sistema de Gestão de Base de Dados |
| SO | Sistema Operacional |
| TDQM | Total Data Quality Management |
| TIQM | Total Information Quality Management |
| TQM | Total Quality Management |

# Capítulo 1

# Introdução

## 1.1 Sistemas de suporte à decisão

A intervenção humana no campo social e organizativo é por natureza subjectiva. Indubitavelmente, encontramos na génese da tomada de decisão uma realidade egocêntrica e singular, ao sabor de sentimentos e suspeições. Esta permanente ilusão tem conduzido algumas organizações a tremendos sucessos, mas muitas outras a irremediáveis infortúnios. Geralmente, o exercício de tomada de decisão é confrontado pela confluência de factores diversos e muitas vezes, antagónicos entre si. Na tomada de decisão, a subjectividade e a presença do risco destacam-se como características predominantes. A introdução de tecnologias, nomeadamente aquelas que sustentam as actividades de gestão estratégica das organizações, tem procurado contribuir para o decréscimo dessa subjectividade e do risco associado à decisão, os *Sistemas de Data Warehousing* (SDWs) são disso um exemplo. A realidade tecnológica tem percorrido um curto, mas incisivo e determinante caminho no seio da sociedade, em geral e das organizações, em particular. Hoje em dia, a diferenciação entre organizações não está somente determinada pela introdução ou uso da tecnologia na condução das suas actividades e na produção dos seus bens e serviços. Cada vez mais, a diferenciação estratégica é resultado das opções tomadas em momentos chave, em especial, no aproveitamento de oportunidades de mercado, na exploração dos pontos fracos da concorrência e na conversão ou camuflagem das debilidades internas. Logo, o sucesso das organizações, num mercado de apetite voraz e deambulante no seu trilho, não se compadece com situações dúbias, desarranjos internos ou más decisões.

Os SDWs vieram adoptar um papel de relevo no domínio dos processos de tomada de decisão, assumindo claramente a vanguarda, relativamente aos sistemas de suporte à decisão anteriormente disponíveis, como uma plataforma tecnologicamente capaz de disponibilizar um conjunto de meios potenciadores de ir ao encontro das preocupações, necessidades e ambições mais essenciais reveladas pelas organizações, em particular pelos seus agentes de decisão. Os SDWs são uma ferramenta muito útil no suporte às actividades quotidianas dos agentes de decisão, essencialmente, na condução táctica das organizações. Este auxílio revela-se importante, na medida em que torna o processo de tomada decisão mais rápido e efectivo, flexibilizando o acesso a mais dados, melhores e mais bem organizados, e garantindo uma maior confiança aos agentes sobre a credibilidade da informação que disponibiliza. Estes são alguns dos vectores basilares que justificam a integração e a exploração de SDWs no seio das organizações. Todavia, a sua construção não corresponde, só por si, ao sucesso da sua implementação – os casos de insucesso são muitos, conforme se pode comprovar através de diversos estudos sobre esta realidade.

As razões para o insucesso dos SDWs correspondem muitas vezes, e paradoxalmente, à escassa ou inexistente obtenção das vantagens enunciadas, justificada, na maioria dos casos, por:

- Deficiências no planeamento do projecto e na implementação do sistema.
- Ineficientes processos de extracção e refrescamento dos dados.
- Desadequadas estruturas tecnológicas de suporte à operacionalidade do sistema.
- Incorrecta modelação dimensional dos dados existentes.
- Falhanços graves ao nível do levantamento de requisitos dos utilizadores.

A este rol de pressupostos, poderemos ainda adicionar alguns factores relacionados com a qualidade dos dados processados pelo próprio sistema. Esta questão, tantas vezes negligenciada pelas organizações, tem vindo a assumir nos últimos tempos um relativo protagonismo, revelando-se cada vez mais como uma fonte da origem desses eventuais fracassos (ou razoáveis sucessos) no desenvolvimento desses sistemas.

As preocupações com os dados têm-se mostrado um assunto obscuro nas organizações, pouco visível, não encarado como um tema transversal a toda a organização e dissimulado, voluntária ou involuntariamente, por outros problemas. Normalmente, a perda de desempenho é associada à ineficácia das ferramentas de interrogação ou a questões de *hardware* (e.g. pouca capacidade de memória ou fraco cumprimento do processador). Porém, os verdadeiros motivos podem residir em questões relacionadas com os dados. A permanência no sistema, dos designados dados dormen-

tes[1] [Inmon et al., 1998], nunca ou raramente utilizados, associada a elevadas taxas de crescimento dos repositórios de *Data Warehouse* (DW) (em [Kimball & Caserta, 2004] referem-se algumas tabelas monstruosas com 200 milhões de linhas e apresentando taxas de crescimento anual na ordem dos 50%), degrada, certamente, os tempos de resposta na entrega dos dados aos consumidores finais. Os problemas dos dados tendem a ser dissimulados voluntariamente porque as consequências visíveis para muitos responsáveis organizacionais são aparentemente mínimas, quando comparadas com as falhas concretas do sistema. Por outras palavras, uma falha de sistema obriga a repensar o sistema existente e a ponderar a sua troca ou actualização. Quanto aos dados, enquanto não se reconhecerem e especificarem os verdadeiros custos associados à sua pouca qualidade, as organizações tendem a menosprezá-los e a deixá-los no fim da lista de prioridades. A qualidade dos dados não tem sido assumida como a grande responsável pela falha de muitos SDWs, quando são eles a razão da existência destes sistemas. Felizmente, as tendências recentes mostram uma reviravolta neste assunto. Possivelmente, porque os problemas de desempenho permanecem, as quebras de sistema subsistem, o armazenamento dos dados continua a crescer a ritmos estonteantes e os dados são assumidos, cada vez mais, como o instrumento de resposta diferenciador num mercado agressivo. Esta realidade pode ser constatada no mais recente estudo da *Pricewaterhousecoopers* [Kenyon et al., 2004]. A qualidade dos dados em SDWs pode ser analogamente percepcionada como a qualidade do sangue no sistema circulatório dos seres humanos.

Ora, actualmente, o problema coloca-se a nível da implementação de eficazes políticas promotoras da elevada qualidade dos dados nos SDWs. Ao contrário dos normais incrementos tecnológicos, capazes de produzirem efeitos apenas pela sua concretização, as questões referentes aos dados não resultam de simples adendas de *software* e *hardware*, nem são tangíveis numa única execução. Antes, implicam uma rotina continuada de procedimentos de tratamento destas questões e um conjunto de meios técnicos e metodológicos que assegurem o cumprimento funcional nas organizações. Esta questão é naturalmente importante na medida em que os SDWs determinam a presença de dados de boa qualidade, situação que faz prever rectificações a montante do DW, isto é, no *Sistema Operacional* (SO) e assim reverter o princípio *garbage in, garbage out* em *quality in, quality out*.

Na realidade, os SDWs apresentam características muito próprias no domínio dos sistemas de informação, porque se situam no mais alto plano da acção estratégica das organizações, desmembrando-se do seu plano estritamente operacional e logístico. A interacção deste tipo de sis-

[1] Tradução do inglês: *dormant data*

tema com os seus utilizadores desenvolve-se apenas com os meios humanos norteadores da condução do negócio – os agentes de decisão – e é realizada apenas com o motivo de lhes dar informações. Os utilizadores apresentam-se, assim, como meros consumidores de informação e conhecimento. Estes interagem com o sistema, geralmente, através de ferramentas que potenciam o processamento analítico dos dados *On-Line Analytic Processing* (OLAP). Do ponto de vista técnico, os SDWs podem assumir, igualmente, uma postura de concentrador dos dados dispersos pela organização permitindo a integração e homogeneização destes. Assim, dispõem-se como um meio único capaz de fornecer correcta e adequadamente os dados da organização.

Neste contexto, o desenrolar das actividades dos agentes de decisão assenta, primordialmente, nos resultados oferecidos pelos SDWs. A exigência de resultados que apresentem propriedades inequívocas, no que respeita à correcção, relevância, acessibilidade, completude, segurança, frescura e consistência das informações, balizam uma área de conhecimento específica. É sobre esta área de conhecimento que se pretende desenvolver a presente dissertação.

## 1.2 Qualidade dos dados e SDWs

O resultado da introdução tecnológica no seio organizacional, ao longo das últimas décadas, tem produzido vastíssimos campos férteis de dados, que se apresentam como um filão riquíssimo por explorar. A tradução deste valor potencial em realidade, apenas poderá ocorrer quando os dados forem assumidos como uma fonte incontornável na criação de vantagens estratégicas. O mercado actual acarreta o dinamismo das acções do negócio e implica, entre outros: o manuseamento de dados de origem interna e externa à organização; o cruzamento destes pelos diversos decisores; a perda de controlo directo das fontes de informação pelos decisores; a prontidão da resposta à concorrência ou no caminho a seguir; a sintonia entre os indicadores organizacionais gerais e a pormenorização dos dados que suportam esses indicadores.

O SO organizacional não responde adequadamente perante estas exigências e como tal não satisfaz os desejos dos consumidores. A principal razão prende-se com a natureza intrínseca destes sistemas, pois encontram-se orientados para a organização e armazenamento dos dados [Kimball et al., 1998]. A sua estrutura interna apresenta dificuldades impeditivas de manipulação eficaz e eficiente dos enormes volumes de dados armazenados nas organizações e, normalmente, necessários para a tomada de decisão. Os SDWs contribuem precisamente como forma de ultrapassar estas dificuldades porque assentam numa estrutura não-volátil e integrada dos dados. Deste modo, possibilitam a permissão de interrogações complexas orientadas aos assuntos organizacionais, dirigidas sobre planos temporais alargados e agindo de modo integrado [Inmon, 1996].

Os dados não se podem apresentar aos decisores desfasados do seu propósito ao uso. A realidade organizacional e especificamente, a relativa aos SDWs não se deve associar com dados feridos nas suas características e não geradores de mais valias. Mas, acontece. Os SDWs encontram-se polvilhados por dados fracos ou doentes, monos e de valor nulo ou residual. Diversas investigações têm revelado esta realidade e condicionam o sucesso dos SDWs a políticas que relevam os dados aos patamares mais elevados no contexto organizacional. Assim, os SDWs e os dados que os sustentam, devem contribuir, no momento da tomada de decisão, como um factor de natureza objectiva na sua essência e capaz de minimizar o risco e a subjectividade, sem contudo pôr em causa a audácia associada à decisão.

A importância da qualidade dos dados nos SDWs é, naturalmente, crucial e pode determinar o seu sucesso e consequentemente o da própria organização. Assim, explicada a essência de um sistema específico deste tipo, nomeadamente ao nível do seu posicionamento na organização, ao público-alvo ao qual se destina e às potencialidades associadas, parece claro que a fundamentação deste em dados tendencialmente erróneos, que servem de base à informação e conhecimento, significará a produção de resultados inadequados. Doutro modo, podemos inferir que os SDWs são tão bons quanto os dados neles contidos.

## 1.3 Motivação e objectivos da dissertação

Os SDWs apresentam propriedades no domínio do tratamento dos dados com o objectivo de os transformar num recurso de valor estratégico que pode ser explorado e utilizado pelas organizações no seu dia-a-dia. Esta situação verifica-se, especialmente, na análise e interpretação dos dados de forma a gerar nova informação e conhecimento estratégico essencial como suporte ao desenrolar do exercício da tomada de decisão. Neste contexto, o interesse em investigar, e dominar, a temática da qualidade dos dados no seio dos SDWs surge como a principal motivação para o desenvolvimento deste trabalho. A par desta motivação surgem outras em torno da mesma temática, como sejam:

- A confrontação de estudos e casos reais que mostram a fraca qualidade dos dados como um factor para o insucesso dos SDWs.
- O reconhecimento dos SDWs como a plataforma tecnologicamente capaz de sustentar o processamento analítico dos dados.
- A comprovação da influência da qualidade dos dados no sucesso das organizações.
- O surgimento desta área como tema emergente na implementação de DWs.

Este último aspecto faz-se notar, por um lado, pela escassez de trabalhos científicos produzidos sobre esta temática, em especial, originários do território nacional, e por outro lado, na adopção de mecanismos e políticas de gestão dos dados que não vão além das tentativas de automatização de processos de detecção e consequente rectificação dos defeitos ocorridos. Algumas irregularidades não podem ser resolvidas pela simples aplicação tecnológica (e.g. a deficiente captação dos dados vitais de um paciente recolhidos no momento duma intervenção cirúrgica). Os estudos e investigações começam a florescer, constatam a forte correlação entre os dados doentes e os elevados custos e naturalmente, as organizações começam a ficar alerta para esta problemática. O tratamento cabal destas questões envolve outros assuntos, que extravasam o domínio técnico dos SDWs, como sejam:

- A necessidade de mudança da cultura organizacional.
- A adopção da informação divulgada como um produto produzido.
- A tendência para a gestão autónoma dos dados.
- A implementação de políticas preventivas em detrimento de políticas de inspecção e correcção de valores dos dados.
- A focalização na satisfação dos interesses, necessidades e anseios dos consumidores dos dados.
- O melhoramento integral e contínuo dos processos que envolvem os dados, desde as fontes até à apresentação aos consumidores.
- O estabelecimento de políticas que promovam a formação dos consumidores dos dados.

Por estes motivos, e dada a importância que os dados adquirem no contexto actual, como recurso estratégico e gerador de vantagens concorrenciais, a sua gestão tende a deixar de estar sobre a alçada das tecnologias de informação e por conseguinte a assumir uma postura mais independente. Confirmando-se como uma área autónoma e balizada por assuntos específicos que merecem tratamento adequado. Diferentes estudos apontam esta tendência e anunciam o desabrochar desta nova realidade. Portanto, este tema trata-se de um ponto quente no contexto organizacional actual e por isso, deriva numa motivação adicional sobre o assunto em causa.

O principal objectivo da presente dissertação consiste em estabelecer uma plataforma para um sistema de gestão da qualidade dos dados em ambientes de DWs, de modo a garantir a boa qualidade das propriedades dos dados e a disponibilizar os resultados adequados aos decisores. Esta plataforma assenta na integração de princípios metodológicos, técnicas, modelos, estratégias e

instrumentos específicos de identificação e remoção dos defeitos nos dados. Nas organizações, revela-se importante a posse de informações de elevada qualidade com o objectivo de agir no momento da tomada de decisões, bem como, o domínio dos melhores meios tecnológicos capazes de disponibilizar essas informações em tempo útil, de forma eficaz e eficiente. Assim, torna-se possível reconhecer em SDWs, o domínio da qualidade dos dados, como de capital importância de modo a permitir apresentar informações e conhecimentos como recurso com valor estratégico.

Na prossecução da concretização do objectivo geral apresentado, são enunciados de seguida, um conjunto de objectivos de carácter mais específico. Nesse sentido, pretende-se também com a realização desta dissertação:

- Constatar a associação da fraca qualidade dos dados aos insucessos dos SDWs.
- Afirmar a qualidade dos dados como assunto transversal a toda arquitectura dos SDWs.
- Identificar as anomalias dos dados e os processos de tratamento e limpeza associados.
- Identificar as causas e as consequências da fraca qualidade dos dados.
- Dominar as diferentes dimensões que compõem o conceito da qualidade dos dados.
- Reconhecer os dados como um assunto organizacional.
- Identificar e analisar propostas e métodos que visem a melhoria dos dados.
- Confirmar os SDWs como uma estrutura tecnológica adequada para a concentração e integração dos dados das organizações.
- Promover uma gestão autónoma dos dados.
- Confirmar os SDWs como instrumentos estratégicos das organizações.

A enumeração dos objectivos enunciados não pretende abordar toda a problemática em estudo. Conforme iremos verificar, essa realidade é significativamente mais densa e complexa. Contudo, a realização da presente dissertação resultou duma assumpção consciente de não podermos abordar outros aspectos, para além dos referidos e que foram igualmente alvo de reflexão, porque ultrapassam o âmbito deste estudo. Por esse motivo, este trabalho circunscrever-se-á de uma forma objectiva às múltiplas temáticas enunciadas. Assim, o estudo dos diversos pressupostos teóricos tem aqui validade pela transversalidade que consubstanciam nos SDWs organizacionais e que servem os objectivos do trabalho proposto. Pretendemos com esta dissertação cooperar, nos meios académicos e profissionais, tendo em vista um conhecimento mais profundo da realidade relacionada com a qualidade dos dados em ambientes de DW.

## 1.4 Estrutura da dissertação

Tendo em vista responder aos problemas mencionados anteriormente, esta dissertação justifica uma organização que descreve a problemática da qualidade dos dados, partindo de um contexto de âmbito generalista para uma abordagem orientada ao domínio dos SDWs. A realidade actual, predominada pela sociedade da informação ou do conhecimento, tende a mostrar a qualidade dos dados como um assunto quotidiano e um desafio a enfrentar pelos diversos actores sociais. Em ambientes de DW, os meios tecnológicos tradicionais, por si só, mostram um alcance curto na resolução dos defeitos verificados nos dados. Uma outra componente, baseada em princípios metodológicos e que fogem ao alcance técnico dos SDWs tem necessariamente de ser enquadrada, conforme é referido em algumas publicações recentes [Adelman et al., 2005] [English, 2002b]. Esta tendência não deve ser entendida como uma relativização ou menosprezo pela vertente tecnológica que envolve os SDWs e que a orientação desta dissertação exige. Pelo contrário, baseados na importância da plataforma tecnológica de DWs no contexto organizacional, procuramos salientar as questões dos dados, como desafios a serem tratados como assuntos transversais à orgânica das organizações e que a sua resolução apenas é, devidamente, conseguida se estes forem assumidos pela cúpula organizativa. Nesse sentido, a presente dissertação encontra-se estruturada em torno de seis capítulos. Além do presente capítulo, o segundo capítulo descreve a problemática da qualidade dos dados no seio da sociedade, em geral e das organizações, em particular. Assim, são expostos alguns conceitos gerais sobre o tema e é abordado o impacto associado à qualidade dos dados. Em seguida, são apresentados alguns motivos determinantes para a fraca qualidade dos dados, bem como as tendências e os mecanismos sociais de controlo da qualidade destes.

A abordagem da problemática da qualidade dos dados focalizada aos SDWs é relatada no terceiro capítulo. Assim, é perspectivado o embate, em termos de custos e benefícios, relacionado com a qualidade dos dados constantes num DW. Depois, são definidas as tipologias de problemas e identificadas as razões da presença desses problemas nos diversos patamares da arquitectura dos SDWs. Adicionalmente, são discutidas as limitações da tradicional *Área de Retenção dos Dados* (ARD) face aos padrões de exigência organizacionais actuais. Além disso, são apresentados os custos associados à manutenção da qualidade nos dados.

Seguidamente, o quarto capítulo procura estabelecer uma plataforma de um sistema de gestão da qualidade dos dados para SDWs, que assuma uma política preventiva e promova a melhoria contínua dos dados existentes. Neste sentido, um DW é encarado como *Produto-Informação* (PI) resultante da conjugação de meios e matérias-primas (dados), tendo em vista a satisfação dos dese-

jos dos utilizadores finais. A plataforma assenta na interligação entre os instrumentos necessários para a resolução das irregularidades dos dados, ao longo das diferentes camadas dos SDWs, e nos princípios metodológicos propostos pela *Total Data Quality Management* (TDQM), como metodologia de aplicação reconhecida. É igualmente referida a importância duma gestão simultânea dos metadados como componente auxiliar das actividades decorrentes e a administração autónoma dos dados.

No capítulo cinco, é prestada especial atenção aos sistemas de medição dos níveis de qualidade dos dados, como parte integrante da plataforma anteriormente apresentada. As métricas, sobre a qualidade dos dados, procuram captar valores indicativos do cumprimento dos critérios de qualidade predefinidos e desse modo aferir o sucesso dos SDWs. Assim, são apresentadas categorias de métricas e discutidos modelos de combinação e tratamento dos valores obtidos, em vista revelar os objectos (processos, repositórios de dados) prioritários para melhoramento. Além disso, é definido um conjunto de métricas, assentes na orientação do comprovado modelo *Goal Question Metric* (GQM), que apontam para as dimensões dos dados basilares para a garantia de uma boa qualidade dos dados.

Um estudo de caso é exposto no capítulo seis, tendo em vista mostrar a emergência da problemática da qualidade dos dados no contexto organizacional. Assim, é realizada uma análise sobre um objecto concreto do SDW duma organização real. O objecto alvo de diagnóstico respeita aos dados constantes num *Data Mart* (DM) relativo às vendas dos artigos por dia. O objectivo desta análise consiste em observar o nível de maturidade da organização na gestão dos dados e de acordo com esse nível, estabelecer um conjunto de recomendações visando a promoção contínua da elevada qualidade dos dados constantes no repositório. Este estudo foi estruturado em duas etapas distintas. A primeira etapa descreve e diagnostica a qualidade dos dados constantes no DM baseada num conjunto de critérios estabelecidos. Posteriormente, na etapa seguinte, procede-se à enunciação de um conjunto de recomendações assentes na plataforma do sistema de gestão da qualidade em SDWs proposto nesta dissertação.

Finalmente, no sétimo capitulo são descritas as conclusões finais e as principais contribuições e limitações da presente dissertação para a compreensão da problemática relativa à qualidade dos dados em SDWs, bem como possíveis trabalhos futuros, baseados nos assuntos aqui abordados.

# Capítulo 2

# A Problemática da Qualidade dos Dados

O ambiente altamente competitivo em que as organizações se encontram inseridas nos dias de hoje, leva a que estas tentem responder o mais adequadamente possível, por vezes de forma agressiva, face à sua concorrência e como forma de criar vantagens competitivas. A informação assume-se como um aspecto fundamental na vivência das organizações e serve de "arma de arremesso" num mercado cada vez mais exigente e devastador. Esta contingência obriga as organizações a manterem os seus dados em perfeitas condições de qualidade, pois só desse modo a criação de vantagens competitivas baseadas na informação se torna uma realidade.

A qualidade dos dados, tema frequentemente negligenciado na vida das organizações, tem vindo a assumir especial interesse por parte dos investigadores conforme comprovam diversos estudos. O estádio de desenvolvimento das infra-estruturas tecnológicas tem apresentado um crescimento exponencial, mas o alinhamento destes meios com a estratégia organizacional é escasso e fica, muitas vezes, aquém do desejado. A constatação que a componente fundamental dos sistemas de informação, são os dados neles contidos, impulsiona as organizações na busca da garantia de uma qualidade dos dados que sirva de veículo no cumprimento da estratégia da organização.

## 2.1 Os dados como recursos estratégicos das organizações

O domínio da informação pelas organizações, quer públicas quer privadas, assume-se nos nossos dias como o principal recurso económico [Serrano & Filho, 2003] [Eckerson, 2002]. As organizações actuam num teatro de operações, marcado pela globalização e imprevisibilidade do rumo seguido. Logo, gera nestas uma necessidade permanente e progressiva em esgrimir argumentos capazes de possibilitarem, num primeiro momento, a sua manutenção em cena e num momento

posterior, o desenvolvimento de vantagens competitivas que corroam o *status quo* vigente e obriguem a adaptação e mudança nas suas congéneres. A informação assume-se como um factor estruturante das organizações e um utensílio de gestão da organização [Amaral & Varajão, 2000].

Cada vez mais, a certeza em assumir a informação como um importante recurso estratégico na vida das organizações e na sociedade em geral, não tem correspondido a uma transparência na sua definição, e por isso, motiva algumas definições significativamente diferentes entre si [Varajão, 1998]. Associado ao conceito informação, outros dois conceitos encontram-se intimamente interrelacionados: dados e conhecimento. Os dados podem ser definidos como representações de factos sobre as coisas que isoladamente não possuem qualquer utilidade [English, 1999] [Varajão, 1998]. Os dados são "factos ou eventos, imagens ou sons que podem ser pertinentes ou úteis para o desempenho duma tarefa, mas que por si só não conduzem à compreensão desse facto ou dessa situação" [Rascão, 2001].

Enquanto, a informação é definida como "os dados cuja forma e conteúdo são apropriados para uma utilização particular, ou seja, são os dados úteis que permitem a tomada de decisões e que está relacionado a algo que nos faz sentido e nos ajuda a compreender os factos e os eventos" [Rascão, 2001]. A informação corresponde aos dados processados e utilizáveis para auxiliar as tarefas em mãos, assumindo os dados, o papel de matéria-prima para a obtenção do produto final: a informação. Por sua vez a informação é transformada ou vertida em conhecimento. É a partir do conhecimento que as organizações avaliam novas situações, aprendem e geram a mudança [Serrano & Filho, 2003]. O conhecimento é definido como "a combinação de instintos, ideias, regras e procedimentos que guiam as acções e as decisões" [Rascão, 2001].

Após a exposição destes conceitos e identificadas as diferenças entre eles, é o momento do devido esclarecimento sobre a indiferença no uso dos termos dados e informação no decorrer da presente dissertação. Assim, adoptaremos uma visão mais abstracta sobre o próprio conceito de dados. O *output* produzido (informação) a partir de determinados dados, poderá servir como *input* para a obtenção de novo *output*. Partindo deste pressuposto recorreremos à visão de Redman, em que os dados são os factos e figuras associados aos clientes, produtos e serviços, mercado e performance financeira, ou seja, todos os aspectos da vida na era da informação [Redman, 2004]. Estes dados são usados para conduzirem todas as operações da organização porque se assumem como o *input* fulcral para o exercício de tomada de decisão e planeamento.

Ainda sobre os dados e contextualizando ao tema desta dissertação, interessa visualizarmos os dados tendo por base a frequência das suas actualizações. Assim, segundo o estudo [Bouzeghoub & Peralta, 2004], podemos identificar três categorias:

- Dados estáveis: são os dados improváveis de se alterarem (e.g. publicações científicas, nomes de pessoas e países).
- Dados que mudam a longo prazo: são os dados que têm uma frequência de mudança muito baixa (e.g. moradas dos funcionários e moedas dos países).
- Dados que mudam frequentemente: são os dados sujeitos a mudanças intensivas, com uma frequência definida ou de modo aleatório (e.g. informações de trânsito, sinais vitais de um paciente e sensores de temperatura).

## 2.2 Da qualidade

A qualidade é um conceito de reconhecida importância e merecedor de especial atenção no campo organizacional desde a década de cinquenta. *Juran*, *Deming* e *Crosby*, confirmam-se como alguns dos grandes impulsionadores do termo, aqueles que tornaram a qualidade numa ciência e efectuaram a sua aplicação ao campo organizacional. A diversa literatura temática mostra que se trata de um conceito complexo, porque é particularmente abrangente na sua acção, subjectivo no seu significado e progressivo ao longo do tempo. Deste modo, não podemos encarar a qualidade como se de um conceito absoluto se tratasse. Alguns estudos [Helfert & Maur, 2001] [1] recorrem a *Gervin* (1984) no sentido de abordar as diferentes perspectivas de análise sobre o conceito qualidade. Estas perspectivas enfatizam os seguintes domínios: cliente, produto, fabrico, excelência e valor. Assim, corroboram sobre a abrangência e relativismo do termo, inviabilizando, por isso, a centralização apenas num único ponto de vista.

Em meados dos anos setenta, *Juran* define a qualidade como uma medida de adequação ao uso[2], isto é, para que um produto ou serviço possua qualidade, as características do produto ou serviço devem ir ao encontro às expectativas do cliente. Esta definição foi posteriormente adoptada pelas normas ISO [Cordeiro, 2004]. Uma segunda definição, conduzida por *Deming*, em 1982, estabelece que a qualidade representa a melhoria contínua de produtos e processos, visando a satisfação dos clientes. Uma outra definição, avançada por *Crosby*, em 1979, orienta a qualidade como a conformidade de um produto ou serviço com as especificações ou requisitos previamente estabe-

[2] Projecto de investigação participado pela Comunidade Europeia, sob o programa de investigação ESPIRIT IV.

lecidos. Esta noção sofreu uma ligeira alteração em vista aclarar o próprio conceito e assim, *Crosby*, em 1992, passou a defini-la como a conformidade do produto com os requisitos dos clientes [Cordeiro, 2004].

As diversas definições apresentadas expõem uma sintonia de orientação ao revelarem, não apenas, a necessidade em garantir a conformidade projectada, a chamada qualidade de conformidade, mas igualmente, em assegurar as especificações a que se pretenda o produto satisfaça, a chamada qualidade de projecto. Adicionalmente, pode ainda ser considerada a qualidade de uso ou pós venda e que se ocupa de aspectos como a manutenção, a fiabilidade e a disponibilidade [Marques, 1994]. Neste sentido, em [Ganhão, 1994] é defendida a necessidade do sucesso nestas questões para que a qualidade seja uma realidade e acautelada no seu todo, ou seja, se a qualidade de projecto não for atingida, por muito bem que decorram as fases seguintes, não é possível assegurar um produto ou serviço adequado ao uso. Igualmente, se a qualidade de conformidade não for conseguida de modo satisfatório, por muito bom que seja o projecto a adequação ao uso será afectada. Daqui, subentende-se o prolongamento do conceito de qualidade e implicitamente a sua gestão, a todas as fases da vida do produto, desde a sua concepção à sua fabricação e posterior utilização pelo cliente. Ainda no que respeita à qualidade, *English* defende que a qualidade não se trata de um luxo, nem corresponde ao facto de ser o melhor da sua classe e evoca a *Total Quality Management* (TQM), para referir que a qualidade existe apenas na óptica dos consumidores e se baseia na noção de valor, ou seja, a forma que estes percepcionam os produtos em vista a atingir de forma consistente as suas expectativas e anseios [English, 1999].

Na diversa literatura temática [Cordeiro, 2004] [Vassiliadis, 2000] [Ganhão, 1994] [Strong et al., 1997] [Ballou & Tayi, 1998], quer no âmbito da qualidade da gestão organizacional quer no âmbito desta dissertação, observa-se que a solução de *Juran* é a mais aceite e representativa do conceito porque expõe a necessidade em satisfazer os desejos dos clientes no momento da especificação do produto e implicitamente estabelece o cumprimento das especificações projectadas. Ora, a adequação ao uso é determinada pelas características ou propriedades do produto ou do serviço que, ao longo da utilização, o utilizador pode reconhecer como benéficas para ele. Estas características, que compreendem as propriedades e atributos do produto ou serviço, são expressas ou não pelos utilizadores, mas o júri final da adequação ao uso (qualidade) é ele e apenas ele [Ganhão, 1994]. Em [Strong et al., 1997] faz-se referência a *Deming* sobre o facto da avaliação da qualidade não se poder dissociar dos clientes que consomem os produtos. Esta linha de pensamento, conduz a outro estágio do desenvolvimento da qualidade, em que esta deixa de ser encarada do ponto de vista negativo (ausência, dispendiosa ou inesperada), mas antes numa óptica positiva (esperada, desejada e legítima) [Silva, 2003]. Assim, a qualidade deve ser enfrentada

como uma matéria que transcende muito a produção e assume um papel crucial nas opções estratégicas das organizações [Marques, 1994]. Ainda em [Silva, 2003], defende-se que a qualidade se torna mais do que a simples satisfação do cliente, deverá ter como fim último a sua sedução e encantamento.

Em vista a satisfação dos seus anseios, os clientes avaliam os produtos ou serviços com base nas características inerentes ou não nestes. Esta situação permite revelar que as características dos produtos consideradas adequadas para um cliente podem não ser para outro [Ballou & Tayi, 1998]. Com base nesta assumpção o conceito é catapultado para um patamar multidimensional [Wand & Wang, 1996] [Pipino et al., 2002], ou seja, distintos consumidores percepcionam os mesmos produtos e serviços, baseados nas diferentes características ou dimensões, implícitas ou explícitas, inerentes aos mesmos e que visam satisfazer as suas diversas necessidades e desejos. Assim, é pressuposto ter em linha de conta os requisitos dos clientes, no momento de planear o produto. Por exemplo, a aquisição duma garrafa de vinho por parte de um consumidor poderá resultar de um complexo processo de escolha na determinação do vinho mais adequado. Algumas questões merecem ser devidamente enquadradas, como sejam a gastronomia que o vinho vai acompanhar (e.g. prato de carne ou peixe), o simbolismo associado à refeição (e.g. refeição do dia a dia ou véspera de natal), o custo da sua aquisição e determinadas características inerentes ao vinho (e.g. elaboração efectuada com castas de reconhecido valor). Portanto, além das características intrínsecas ao próprio vinho, um conjunto de outras questões necessitam de ser contextualizadas para que o processo culmine com sucesso. A experiência acumulada pelas organizações no campo da qualidade dos produtos e serviços permitiu e serviu de suporte para a transferência do conceito ao domínio dos dados ou das informações, conforme se mostra o estudo [Wand & Wang, 1996] que considera a qualidade dos produtos ou serviços como dependente dos seus processos de concepção e fabrico. Tal e qual, a qualidade dos dados depende dos processos de desenho e produção inerentes à geração dos dados.

## 2.3 Da qualidade nos dados

### 2.3.1 A adopção do conceito pelas ciências informáticas

A gestão da qualidade dos dados é uma preocupação assumida desde meados da década de sessenta por parte dos investigadores em estatística [Scannapieco & Catarci, 2002]. As suas preocupações centravam-se no tratamento de conjuntos de dados (e.g. a duplicação de valores no mesmo conjunto de dados). É, ainda hoje, uma área de ocupação nas investigações no ramo da estatística [Brackstone, 2001]. Durante a década de oitenta, as investigações no campo da gestão

assumiram a problemática da qualidade dos dados como fonte de pensamento e objecto de estudo. Apenas no início da década de noventa, o tema mereceu a devida atenção pelas ciências informáticas. Pois, até essa altura não se registam investigações, estudos ou livros temáticos. Em [English, 2001] perspectiva-se a história e regulação da qualidade dos dados de forma similar à ocorrida durante as grandes eras agrícola e industrial. Assim, a década de noventa é dividida em duas partes iguais. Na primeira metade, dá-se o nascimento da problemática em torno da qualidade dos dados como área merecedora de devida atenção por parte dos investigadores. Nesta fase surgiram as primeiras investigações a nível académico como sejam: a tese de *Mark Hansen*, *Zero Defect Data: Tackling the Corporate Data Quality Problem*, pelo *Massachussets Institute of Technology*; o programa de TDQM, desenvolvido por *Wang* [17] e promovido igualmente pelo mesmo instituto; a elaboração do primeiro livro sobre o tema, por *Redman*; o aparecimento das primeiras tecnologias de limpeza dos dados que visavam debelar algumas irregularidades nos dados e surgiram algumas conferências de ciências informáticas abordando o assunto.

Na segunda metade da década de noventa, deu-se um aumento significativo dos problemas relacionados com a qualidade dos dados, provocado pelo aumento em qualidade e quantidade das tecnologias de *software* e pelo aceleramento na perda do controlo dos processos de gestão dos dados [English, 2001]. O crescimento exponencial da circulação e processamento de dados, em especial, motivada pelo uso da Internet e toda a economia vindoura inerente, acentuou os problemas nos dados existentes e potenciou novos conjuntos de falhas nos dados. A par desta tendência, o domínio da qualidade dos dados passou a ser encarado como área de estudo autónoma e, por isso, diversas investigações, estudos, relatórios e soluções de limpeza de dados acentuaram a sua acção em torno da problemática dos dados. O crescimento acentuado de tecnologias de *software* neste sector, a realização de conferências mundiais atraiu muitos investigadores para este tema, as organizações, tanto públicas como privadas, sentiram a necessidade em possuir dados de elevado grau de qualidade e desassossegaram-se na ânsia de soluções que alcançassem os seus intentos [English, 2001].

Na entrada de milénio assiste-se à tentativa de passagem do domínio da qualidade dos dados a um patamar de amadurecimento da própria ciência dos dados. As desgostosas experiências, causadas pela deficiência da qualidade dos dados, impulsionaram num primeiro momento, as organizações a enveredarem esforços no sentido de tratar estas questões e num momento posterior, a encontrarem-se alerta para as reais vantagens em garantir a melhor qualidade dos dados e por isso, assumem a qualidade dos dados como uma das prioridades a resolver [Eckerson, 2002] [Wang, 2004]. A implementação com sucesso de novas plataformas informáticas, como seja: SDW, *Enterprise Resource Planning* (ERP), *Customer Resouce Management* (CRM) e aplicações

OLAP, não pactua com dados de qualidade inferior. Assiste-se igualmente a iniciativas no campo legislativo que visam regular e orientar esta área, traduzindo a importância que este assunto tem assumido nos últimos tempos [Wang, 2004] [Kyl, 2005] [2].

### 2.3.2 As tentativas de definição de qualidade dos dados

As investigações relativas à qualidade dos dados no campo das ciências informáticas têm multiplicado e abordam um vasto leque de questões inerentes ao tema. Esta proliferação de estudos contribui, assim, para um aumento da dificuldade na definição do conceito. Os estudos desenvolvidos, geralmente, não apresentam consenso sobre a abrangência e outros aspectos essenciais deste tema. Os motivos justificam-se tanto por conceitos sobrepostos entre si, como por ópticas divergentes sobre os mesmos termos, como seja a terminologia das características ou dimensões dos dados, a este propósito em [Wand & Wang, 1996] é observado que mesmo o termo exactidão possui diferentes sentidos. As investigações na área da qualidade dos dados são altamente interdisciplinares e complementares, por isso, contribuem ainda mais para a dificuldade de definição do conceito [Chung et al., 2002].

Dado o emaranhado de caminhos seguidos pelas diferentes investigações, alguns estudos [Scannapieco & Catarci, 2002] [Rasmussen, 2004] [Wang et al., 1994] tentam expor uma retrospectiva da literatura sobre a qualidade dos dados em torno de três perspectivas: a ontológica ou teórica, a intuitiva e a empírica. Em [Lee et al., 2000b] é acrescentada a perspectiva arquitectural. A perspectiva ontológica define um conjunto de conceitos teóricos baseados na problemática em causa. A perspectiva intuitiva presume a existência de competências e conhecimentos sobre a matéria abordada. Trata-se da constatação das experiências, geralmente negativas, provocadas pela qualidade dos dados. Por fim, a perspectiva empírica pretende adoptar conceitos e práticas empregues noutros âmbitos. É o caso da conceptualização dos dados e da sua qualidade intrínseca como se de produtos comuns se tratassem.

#### A perspectiva ontológica

A perspectiva ontológica concentra-se numa óptica interna, ignorando a análise de requisitos dos consumidores e está orientada para o desenho do sistema e produção dos dados. Assim, é definido um conjunto de assumpções, postulados e definições, de modo a gerar as dimensões da qualidade dos dados [Helfert & Herrmann, 2002]. As dimensões servem de base para o desenho do sistema de informação, através da enunciação de objectivos concretos de qualidade dos dados, permitindo a orientação deste no sentido de reflectir os aspectos do mundo real [Wand & Wang, 1996]. Em [Orr, 1998] é oferecido, igualmente, um ponto de vista teórico sobre a qualidade dos

dados, definindo-a como a medida de aceitação entre as vistas dos dados proporcionadas por um sistema de informação e os mesmos dados no mundo real. Enquadram-se igualmente nesta perspectiva teórica as aproximações que visam apresentar taxionomias sobre as deformidades verificadas nos dados [Kim et al., 2003] [Oliveira et al., 2005a].

## A perspectiva arquitectural

A perspectiva arquitectural transmite a ideia da utilização de meios tecnológicos (SDW, CRM, ERP) capazes de melhorar o nível de qualidade nos dados, pela focalização da sua acção na forma como os dados se apresentam armazenados e disponíveis para partilha. Esta aproximação assenta no projecto de investigação, *The Foundations of Data Warehouse Quality*[3] - *Data Warehouse Quality* (DWQ) e propõe uma plataforma de arquitectura e um repositório de metadados, que descrevam todos os componentes do DW, num conjunto de meta-modelos que são adicionados a um meta-modelo de qualidade. Este último define para cada meta-objecto do DW as correspondentes dimensões e os factores de qualidade relevantes [Vassiliadis et al., 1999]. A investigação pretende centrar-se numa abordagem em que a qualidade dos dados possa ser assegurada numa concepção mais técnica e tem evoluído em vista a apresentar meta-modelos capazes de capturar, igualmente, as componentes dinâmicas do DW. Esta aproximação releva a importância da existência de metadados de boa qualidade para o sucesso de um sistema de gestão da qualidade dos dados [Jarke et al., 2003].

## A perspectiva produto-informação (empírica)

A perspectiva empírica defende uma gestão da qualidade dos dados idêntica à gestão da qualidade dos produtos e serviços convencionais. Assim, a informação como produto deverá resultar da aplicação de processos de fabrico, materiais e serviços ao consumidor [Wang et al., 1998]. A complexidade envolvente à definição da qualidade dos produtos e serviços, e explicada pelos seus aspectos de natureza subjectiva, multidimensional [Wand & Wang, 1996] [Pipino et al., 2002] e volátil, revela-se de maneira similar quando aplicada ao domínio dos dados. Estes, apesar de possuírem determinadas características divergentes aos produtos e serviços comuns, apresentam questões de gestão idênticas, tanto ao nível de processamento, como ao nível das características ou dimensões inerentes. A este pretexto, em [Wang, 1998], considera-se que para aumentar a produtividade, as organizações devem gerir as informações como gerem os produtos, devendo a qualidade associada ao produto ser aplicada à informação para a obtenção do designado PI. Ainda à luz do mesmo autor, a informação resulta do tratamento e operacionalidade das matérias-

[3] Projecto de investigação participado pela Comunidade Europeia, sob o programa de investigação ESPIRIT IV.

primas (dados), pelos sistemas de informação (processos). Em [English, 1999] é percepcionada de modo análogo a visão da qualidade dos dados como se da qualidade de um produto se tratasse. Ainda segundo *English*, os dados são detentores de características próprias que satisfazem ou não os consumidores e possuem, também, processos próprios na criação, manutenção e utilização idênticos aos produtos convencionais. Em [Strong et al., 1997], esta linha de pensamento é reforçada, focalizando o ponto de vista do consumidor sobre os dados. A qualidade dos dados é assumida como a aptidão para o uso por parte dos consumidores, para isso, é visualizada a produção e o armazenamento dos dados como um sistema de produção de dados, interagindo com os diversos intervenientes (produtores, administradores e consumidores dos dados). Em [Redman, 2004], é referido que os dados de alta qualidade resultam de processos bem definidos e geridos, que criam, armazenam, movem, manipulam, processam e usam adequadamente os dados. Porém, alerta para algumas divergências fulcrais entre os dados e os produtos convencionais. O afastamento entre algumas características dos dados e os produtos e serviços é, igualmente, observado em [Wang, 1998] (e.g. a reutilização dos dados na produção de novas informações).

### 2.3.3 A multidimensionalidade dos dados

As aproximações, referidas anteriormente, vêm corroborar o estudo [Ballou et al., 2004], que aponta para a maioria das investigações centrarem-se na busca do nível de qualidade adequado na perspectiva do utilizador, ou seja, baseiam-se no princípio de *fitness for use* ou atingir as expectativas do utilizador final. Em [Ballou & Tayi, 1998] é enfatizada a relativização e multidimensionalidade do conceito de qualidade dos dados. A qualidade de determinados dados pode ser considerada apropriada para os anseios dum consumidor, ao passo que pode não ser suficiente para outro consumidor. A definição do nível apropriado da qualidade dos dados está dependente do seu contexto [Pipino et al., 2002]. A aprovação da perspectiva multidimensional, associada à qualidade dos dados, surge igualmente por uma iniciativa legislativa, promovida pelo *Office of Management and Budget* (OMB). Esta entidade define qualidade dos dados como os dados que oferecem utilidade, objectividade e integridade aos consumidores de informação [Kyl, 2005] [2]. Em [Olson, 2003], a qualidade dos dados depende tanto das utilizações pretendidas (óptica percepcional) como dos dados em si mesmos (óptica factual). Ora, este pressuposto configura a constituição dos dados com aspectos de natureza factual e percepcional. A visão factual ou imparcial dos dados permite aferir quantitativamente da qualidade dos mesmos, ou seja, possibilita a avaliação da qualidade dos dados independentemente de outros factores, como sejam os utilizadores ou a decisão a tomar. A visão percepcional ou contextual faz depender a qualidade dos dados da utilidade que estes possuem na satisfação das necessidades dos utilizadores ou da decisão a tomar [Skri-

letz, 2002] [Shankaranarayan, 2005]. Em vista a satisfação das utilizações pretendidas, os dados devem respeitar de forma, eficaz e eficientemente, um conjunto de dimensões que garantam a qualidade dos dados para cada caso concreto. A qualidade dos dados deverá envolver os dados certos e correctos no local certo para o consumidor completar a tarefa em mãos [Redman, 2004]. Enquanto que no estudo [Wand & Wang, 1996], é defendida uma noção de qualidade dos dados dependente do uso corrente dos dados. Em decorrência, a qualidade dos dados pode expressar-se numa hierarquia de diferentes categorias de características, que são posteriormente refinadas em dimensões de qualidade dos dados [Jarke et al., 2003]. Assim, é possível ser definida como um valor agregado sobre um conjunto de critérios de qualidade, que indique quanto bons são os dados que possuímos, tendo em conta as exigências do negócio [Müller & Freytag, 2002]. Em suma, a adopção de critérios de medida dos dados tenta transmitir de modo transparente e objectivo a qualidade inerente aos mesmos. Esta posição é partilhada em [Brackett, 1996], que considera uma consistente qualidade dos dados, quando o estado da qualidade dos dados é sobejamente compreendido e conhecido.

Do exposto, verifica-se que as investigações se centram em torno do conceito *fitness for use*, pois assentam em duas traves mestras fundamentais. Por um lado, a exigência de um controlo de qualidade de conformidade dos processos de produção dos dados. Por outro lado, a completa satisfação ou superação das necessidades e desejos dos consumidores dos dados, de forma a estes cumprirem as suas tarefas. Os dados têm qualidade se satisfizerem os requisitos para a sua utilização e verifica-se uma falha da qualidade sempre que não satisfaça um requisito [Olson, 2003].

## 2.4 As dimensões da qualidade dos dados

Conforme referido anteriormente, a natureza multidimensional intrínseca ao conceito de qualidade dos dados aponta para o entendimento sobre as dimensões observadas em seu redor. A assumpção da dimensão exactidão como aglutinadora do próprio conceito de qualidade dos dados tem sido contrariada como demonstram vários estudos [Ballou & Tayi, 1998] [Wang et al., 1994] [Strong et al., 1997] [Pipino et al., 2002]. Estes estudos apresentam a exactidão como condição necessária, mas não suficiente. Neste pressuposto, contrapõem um conjunto de dimensões importantes e necessárias para a concretização do objectivo em obter os melhores dados possíveis. Os objectivos a atingir são determinados, geralmente, pelas exigências dos consumidores e a sua obtenção corresponde à satisfação dos anseios destes, que os convertem em benefícios para as organizações. Portanto, é pelo cumprimento das dimensões dos dados, que as organizações encaram estes como recursos estratégicos potenciadores da criação de vantagens competitivas. Deste modo, estão criadas as condições logísticas sobre os dados, que permitem às organizações

enfrentar ambientes de mercado hostis, quer pela resposta às suas congéneres, quer pelo impulsionamento condicionante sobre o teatro de operações envolvente.

A importância assumida pelas dimensões na gestão da qualidade dos dados conduz a um estudo mais específico, em torno de seis componentes essenciais: a imperfeição dos dados; as categorias das dimensões; a área de domínio das dimensões; a hierarquia das dimensões; o estabelecimento de associações entre dimensões e o relacionamento entre as dimensões e os diversos intervenientes dos dados. Um outro assunto igualmente importante respeita à captação de indicadores ou medidas sobre a qualidade dos dados, baseadas nas diferentes dimensões. Este tema será debatido em detalhe no capítulo cinco, que respeita à obtenção de métricas relativas à qualidade dos dados, de modo a aferir quanto à validade das informações produzidas.

### 2.4.1 Imperfeição dos dados

Recentemente, a investigação [Ballou et al., 2004] faz notar o carácter evolutivo intrínseco ao termo qualidade dos dados. Se anteriormente, a qualidade dos dados correspondia à garantia da exactidão dos dados, desde o início da década de noventa, pelo incremento da utilização da informação como um recurso estratégico, revelou-se o carácter multifacetado da qualidade dos dados. Assim, são consideradas outras dimensões, como sejam, entre outras: a consistência, a completude, a oportunidade e a compreensão. Naturalmente, esta situação contribui ainda mais para o aumento da complexidade do conceito na garantia da qualidade dos dados porque mantém uma ideia de opacidade na sua definição, ou seja, não exige, de modo subtil, a perfeição dos dados, situação impossível e certamente desnecessária [Ballou et al., 2004]. Deste modo, a aptidão para o uso continua a ser aquela que melhor a define. A presença de defeitos nos dados é aceitável, segundo o estudo [Eckerson, 2002], que considera que os dados não necessitam de estar perfeitos para serem úteis. Este pragmatismo é, igualmente, partilhado em [Orr, 1998], ao ser afirmado que nenhum sistema de informação pode assegurar uma qualidade dos dados de 100%. Ainda na mesma investigação, é referido que o problema real não passa por assegurar uma qualidade dos dados perfeita, mas antes que a qualidade dos dados do sistema de informação seja suficientemente segura, atempada e consistente para que uma organização sobreviva e tome decisões razoáveis [Orr, 1998]. Reforçando esta posição, em [English, 1999] é considerado como principal objectivo da organização a maximização do valor dos seus recursos em vista perseguir a missão da organização. Neste sentido, o objectivo da qualidade dos dados consiste em equipar os consumidores com dados, tratados como recurso estratégico, capazes de permitir a inteligência das organizações e possibilitar a tomada de decisões eficazes e eficientes. Para isso, separa o conceito de qualidade dos dados em duas partes. A primeira, designada por qualidade dos dados

inerente e que consiste na representação real dos dados. Uma segunda noção, a qualidade dos dados pragmática é definida como o grau de utilidade e valor dos dados que suportam os processos da organização, em vista a atingir os objectivos traçados. Neste contexto, os dados armazenados num repositório não possuem valor actual, mas apenas valor potencial.

Portanto, a elevada qualidade dos dados não pode ser entendida como a sua perfeição, essencialmente, por dois factores. Um primeiro factor remete-nos para uma abordagem multidimensional subjacente ao conceito. Os consumidores dos dados possuem distintas percepções sobre a qualidade dos mesmos dados e um mesmo consumidor pode atribuir importâncias distintas aos mesmos dados em diferentes instantes de tempo. Um segundo factor compreende as razões de natureza económico-financeira e de ordem logística. A garantia de uma qualidade dos dados perfeita em todas as frentes (dimensões), mostra-se um objectivo tanto inexequível como desnecessário. Nesse sentido, importa enveredar os esforços para a obtenção de uma qualidade dos dados apropriada ao fim destinado, isto é, que satisfaça (ou exceda) os desejos dos consumidores e os auxiliem no processo de tomada de decisão.

### 2.4.2 Categorias das dimensões

As dimensões que compõem os dados são, por vezes, designadas de atributos ou características ou mesmo critérios e correspondem a adjectivações dos próprios dados. Em virtude da vasta amplitude do número de dimensões, o agrupamento destas por categorias ou níveis permite um tratamento mais directo e consequentemente uma gestão mais manuseável e viável dos próprios dados [Watson et al., 2002]. Em [Strong et al., 1997] são identificadas 179 dimensões da qualidade dos dados, enquanto em [Watson et al., 2002] faz-se referência a alguns estudos que apontam, igualmente, para valores na ordem da centena de dimensões. A satisfação das dimensões associa-se a um elevado grau de volatilidade na sua essência. A aceitação duma característica por um consumidor, num instante de tempo, não significa a manutenção dessa vinculação, num outro instante de tempo. Assim, é, igualmente, admitida a possibilidade de vários consumidores poderem estabelecer diferentes percepções e avaliações sobre os mesmos dados e as características que os revestem. Daqui, resulta que a qualidade dos dados depende não apenas da qualidade inerente aos dados, como também, do contexto em que estes se inserem [Wood, 2002] [Scannapieco & Catarci, 2002]. Em consequência, pode-se considerar um problema ao nível da qualidade dos dados quando se verifica deficiências numa ou mais dimensões [Strong et al., 1997]. Adicionalmente, constatamos que distintas investigações veiculam diferentes abordagens sobre esta problemática. Ora, esta dificuldade em garantir uma plataforma consensual de conceitos origina diferentes sentidos de orientação, distintas percepções sobre os mesmos termos e sobreposição

de definições desses conceitos. Neste cenário, alguns estudos de exposição comparativa têm surgido no sentido de clarificar as diversas correntes [Amaral, 2003] [Scannapieco & Catarci, 2002] [Lee et al., 2001].

É possível constatar na literatura alguns estudos que procuram esclarecer e enumerar as principais abordagens sobre as dimensões da qualidade dos dados. O trabalho [Lee et al., 2001] tenta confrontar a perspectiva académica e a perspectiva dos profissionais sobre as dimensões observadas. Para isso, sumaria as investigações académicas que mais se destacam no domínio das dimensões e apresenta as dimensões mais usadas pelos profissionais. É observável, igualmente, o enquadramento das dimensões pelas respectivas categorias. A tarefa de catalogação das dimensões não se apresenta pacífica na sua realização, porque os resultados dos estudos surgem de diferentes orientações adoptadas por cada investigação. Em [Wang et al., 1994] são identificadas quatro categorias de dimensões: intrínseca, contextual, representativa e acessível. Esta classificação é bastante referenciada na literatura e tem servido de plataforma conceptual para alguns estudos académicos, como seja o promovido pelo DWQ, que a contextualiza ao domínio dos DW [Jarke & Vassiliou, 1997] A tabela 2-1 ilustra algumas das diferentes perspectivas.

| | **Intrínseca** | **Contextual** | **Representativa** | **Acessibilidade** |
|---|---|---|---|---|
| [Wang et al., 1994] | Exacta, credível, reputada e objectiva | Valor acrescentado, relevante, completa, oportuna e quantidade apropriada | Compreensível, interpretável, representação concisa e representação consistente | Acesso, fácil de operar e segura |
| [Jarke & Vassiliou, 1997] | Acreditar: exacta, credível, consistente e completa | Utilidade: relevante, usável, oportuna: (fonte corrente, DW corrente, sem volatilidade) | Interpretação: sintaxe, semântica, controlo da versão, origem, pseudónimos | Acesso: disponibilidade do sistema, disponibilidade de transacções e privilégios |
| [Wand & Wang, 1996] | Correcta e sem ambiguidade | Completa | Significativa | |
| [Olson, 2003] | Exacta e confiança | Oportuna, relevante e completa; | Compreensível | |
| OMB [English, 2003c] | Objectividade: exacta, clara, completa e livre de preconceitos | Útil | | Integridade: protecção |
| [English, 1999] | Inerente: conformidade de definição, completa, validade, correcção com as fontes, correcção com a realidade, exacta, e não redundante | Pragmatismo | | |
| [Lee & Strong, 2003] | Reputação: precisa | Utilidade: relevante e completa<br>Confiança: oportuna | | Usável: acesso |

**Tabela 2-1 – Perspectiva das categorias das dimensões dos dados.**

O estudo [Scannapieco & Catarci, 2002] apresenta algumas investigações académicas primordiais e estabelece as similitudes e diferenças adoptadas entre elas, em termos de identificação e significados das dimensões. As propostas das dimensões dos dados estudadas são vertidas em torno de quatro categorias: a aproximação pela definição das dimensões; a modelação de vistas sobre os dados; a medição das vistas sobre os dados e a dependência do contexto. Um outro estudo [Amaral, 2003] confronta as dimensões mais representadas nas abordagens académicas de dezassete estudos considerados. O estudo confirma a existência de um núcleo restrito e relevante de dimensões nas principais aproximações: exactidão, completude, oportunidade, consistência, confiança, compreensão, relevância, interpretação e actualidade. O estudo apresenta um conjunto de dimensões uniforme e transversal em todas as abordagens académicas e dos profissionais [Lee et al., 2001]. Em estudos recentes [Lee & Strong, 2003, 2004], relaciona-se o conhecimento dos diversos intervenientes e a qualidade dos dados. Os estudos têm contemplado cinco dimensões, apresentadas em [Wang et al., 1994] e consideradas vitais pela generalidade dos consumidores: exactidão, completude, relevância, oportunidade e acessibilidade. Em [Ballou & Tayi, 1998] são identificadas quatro dimensões no domínio da qualidade dos dados: exactidão, completude, consistência e oportunidade. Em [Wand & Wang, 1996] é introduzida uma aproximação que focaliza o desenho e a operação nos sistemas de informação [Helfert & Herrmann, 2002]. As dimensões são determinadas pelo mapeamento entre o mundo real e o sistema de informação [Scannapieco & Catarci, 2002]. Os defeitos ocorridos nas dimensões são observáveis pela comparação entre o mundo real e o sistema de informação. A centralização desta aproximação no sistema de informação negligencia a perspectiva dos consumidores dos dados [Helfert & Herrmann, 2002]. O estudo propõe quatro dimensões possíveis de definir objectivamente: completude, ausência de ambiguidade, significativa e correcção. Em [Wang et al., 1994] são construídas teoricamente as dimensões consideradas do ponto de vista dos consumidores, que identificam a exactidão e a correcção dos dados como as mais importantes. É, igualmente, estabelecido um conjunto de quatro categorias de dimensões compostas por dimensões específicas [Helfert & Herrmann, 2002]. A investigação [Jarke & Vassiliou, 1997] propõe uma orientação da qualidade dos dados em torno do desenho das actividades envolvidas na construção de um DW [Scannapieco & Catarci, 2002]. Assim, desenvolve uma ligação entre os factores de qualidade e os grupos de intervenientes envolvidos no projecto de DW [Helfert & Herrmann, 2002]. Em [English, 1999] é proposta a qualidade dos dados em torno de quatro categorias: definição dos dados, conteúdo dos dados, apresentação dos dados e a arquitectura dos dados.

### 2.4.3 Domínio das dimensões

A exactidão dos dados é considerada pela generalidade das investigações como a dimensão mais importante, mas não possui uma definição clara e consensual [Wand & Wang, 1996]. Em [Brackett, 1996] considera-se que não existe nenhum nível de exactidão predeterminado, cada organização deverá definir o nível de exactidão desejado. Nesta dissertação recorreremos à definição abstracta exposta em [Ballou & Tayi, 1998], em que, a exactidão corresponde ao armazenamento correcto dos factos ou valores do mundo real, ou seja, consiste em possuir valores certos e de confiança [Pipino et al., 2002]. Em [Olson, 2003] defende-se que a obtenção dos dados correctos só é possível pela garantia da posse dos valores dos dados certos e pela sua representação consistente e insusceptível de gerar ambiguidades. Assim, é observado que num conjunto de dados constam, além dos dados correctos, os dados incorrectos e que são originados por: representações erradas (e.g. Lx. e Lisboa), valores errados (e.g. o tempo de serviço de um professor é de 170 anos), valores inválidos (e.g. idade = 'FH') e a ausência de valores (figura 2-1).

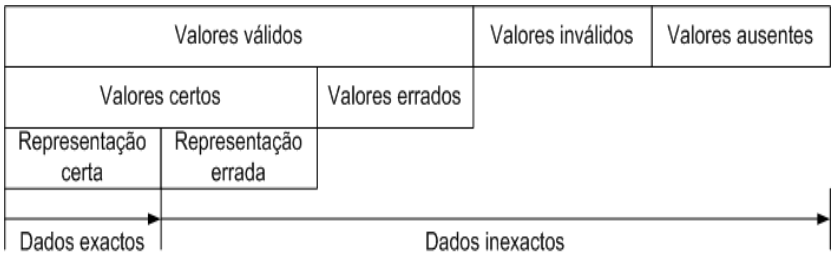



**Figura 2-1 – Constituição de um conjunto de dados [Olson, 2003].**

Num outro estudo [Müller & Freytag, 2002], a exactidão é observada segundo três ópticas: sintáctica, de cobertura e semântica. A primeira descreve as características dos formatos e valores usados na representação das entidades do mundo real. Esta classe compreende a definição dos modelos léxicos (e.g. estruturas de dados) e do formato do domínio dos dados (e.g. unidades de medida diferentes). A exactidão de cobertura refere-se ao controlo dos valores ausentes das colunas ou linhas. Estas situações podem resultar em omissões sobre as entidades existentes no mundo real. Por último, no que respeita à exactidão semântica, o estudo observa a necessidade dos dados serem perceptíveis e não redundantes. Nesse sentido, devem ser satisfeitas, por um lado, as regras de integridade dos dados, que podem ser compostas em: integridade do valor dos dados, integridade das estruturas de dados, integridade da retenção dos dados e a integridade de derivação dos dados [Brackett, 1996] (tabela 2-2). Por outro lado, devem ser efectuados esforços em vista a redução de dados redundantes (e.g. linhas duplicadas) e valores inválidos, que não representam fielmente as entidades do mundo real.

| Integridade | Componente | Exemplo |
|---|---|---|
| Valor dos dados | Domínio de valores dos dados | A data inscrição de um aluno não pode ser superior à data do momento. |
| | Integridade do valor dos dados condicional | O preenchimento do nome do aluno obrigatório. |
| | Os valores por defeito | O campo nacionalidade poderá conter por defeito 'portuguesa'. |
| Estruturas de dados | Integridade referencial | Um aluno só se pode inscrever num exame que exista. |
| | Integridade da estrutura de dados condicional | Um aluno pode estar inscrito em zero, um ou mais exames. |
| Retenção dos dados | | Prevenção nas eliminações dos dados. |
| Derivação dos dados | | Idade do aluno = ano actual – ano de nascimento. |

**Tabela 2-2 – Regras de integridade dos dados [Brackett, 1996].**

Em suma, verifica-se que os assuntos envolventes a esta dimensão mostram-se dispersos e quando associados à multidimensionalidade do conceito de qualidade dos dados, resultam no aparecimento de outras dimensões. Assim, a perspectiva da exactidão como dimensão centralizadora dos diversos referentes aos dados tem dado lugar a outros assuntos, igualmente importantes, no respeita à qualidade dos dados e que se consubstanciam em outras dimensões.

A dimensão completude diz respeito ao armazenamento de todos os dados considerados importantes, de modo a que a ausência ou a insuficiência de detalhe dos dados não seja detectada durante a tomada de decisão [Pipino et al., 2002] [Lee & Strong, 2003]. Indica o grau de presença dos valores numa colecção de dados [Fischer & Kingma, 2001]. A dimensão oportunidade corresponde à exigência dos dados estarem suficientemente actualizados para a execução das tarefas a tratar [Lee & Strong, 2003]. Os dados devem estar disponíveis atempadamente para influenciarem o processo de tomada de decisão [Fischer & Kingma, 2001]. A definição do tempo de apresentação dos dados não pode ser encarado como universal para todas as decisões a tomar, deve depender das circunstâncias e necessidades sentidas pelo executor da decisão (e.g. a necessidade de dados vitais frescos, actualizados a cada 5 segundos, sobre um paciente numa operação não se verifica na execução de um plano de marketing de um produto). Doutro modo, podemos considerar que a oportunidade dos dados consiste no tempo dispendido desde que os dados são gerados pelos sistemas iniciais até ao momento que se encontram disponíveis para os utilizadores [Inmon et al., 1998]. Em [Bouzeghoub & Peralta, 2004] esta questão é tratada envolvendo um nível hierárquico superior: a frescura dos dados disponibilizados. Assim, a frescura dos dados compreende duas dimensões: a oportunidade e a actualidade dos dados. A oportunidade descreve a idade dos dados no sistema. Enquanto, a actualidade consiste no deslocamento temporal dos dados em relação às fontes. A dimensão relevância refere-se à aplicação e utilidade que os dados possuem em vista a execução das tarefas em mãos [Pipino et al., 2002]. Confere o grau de perti-

nência dos dados no processo de tomada de decisão [Brackstone, 1999]. Assim, os dados relevantes podem ser usados para solucionar alguns problemas das organizações [Fischer & Kingma, 2001]. A dimensão acessibilidade responde pela disponibilização dos dados aos utilizadores, ou seja, resume-se à facilidade e rapidez que esses dados podem ser obtidos pelos utilizadores [Pipino et al., 2002]. A dimensão relativa à credibilidade define o grau de confiança que é depositado nos dados divulgados. A dimensão credibilidade compõe-se pela confiança nas fontes de origem dos dados e nos processos de transformação e tratamento dos dados. A dimensão consistência refere-se às anomalias sintácticas e com as contradições entre os dados [Müller & Freytag, 2002]. A interpretação dos dados consiste na disponibilidade de informação suplementar e metadados sobre os dados em vista a sua utilização e representação adequada [Brackstone, 1999].

### 2.4.4 Hierarquias das dimensões

Um outro aspecto, decorrente do assunto anterior, respeita à hierarquia sobre as diversas dimensões. Geralmente, as dimensões resultam da agregação entre dimensões ou servem de *input* para a obtenção de novas dimensões. A definição de uma estrutura hierárquica entre as dimensões não é assunto pacífico capaz de ser desenvolvido sem a contextualização adequada de um cenário real. Em [Müller & Freytag, 2002] é apresentada uma possível hierarquia sobre as dimensões dos dados de modo a que a qualidade destes possa ser definida como um valor agregado sobre um conjunto de critérios de qualidade (figura 2-2).

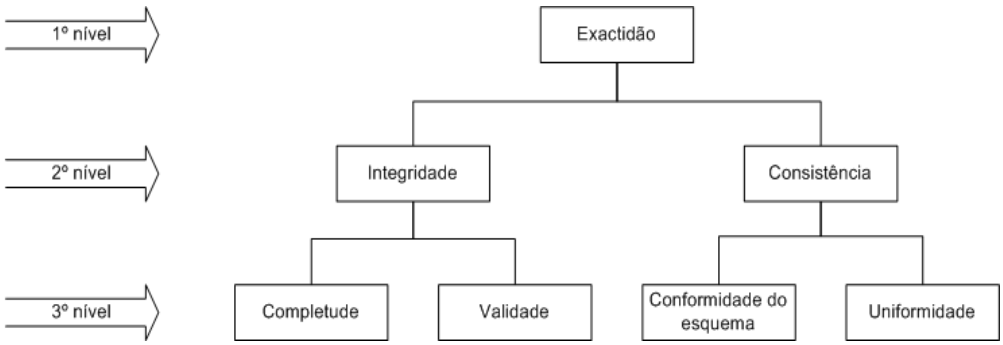



**Figura 2-2 – Hierarquia das dimensões dos dados [Müller & Freytag, 2002].**

Também em [Brackstone, 1999, 2001] se estabelece, no campo estatístico, uma hierarquia entre as dimensões (figura 2-3). O estudo justifica a relevância como a dimensão no topo da pirâmide hierárquica porque a posse de dados perfeitos baseados sobre tópicos irrelevantes conduz a dados inúteis. Assim, obtida a relevância, sem a oportunidade e acessibilidade dos dados, estes não se encontram disponíveis quando necessário. Por último, só após a concretização destas dimensões é que faz sentido centrarmo-nos na exactidão, interpretação e coerência dos dados.


**Figura 2-3 – Hierarquia das dimensões dos dados [Brackstone, 2001].**

### 2.4.5 Associação entre dimensões

O relacionamento entre as diferentes dimensões, que compõem os critérios de qualidade predefinidos no desenho dos sistemas, trata-se de um desafio de interesse crucial. É através da sua correcta definição que a qualidade dos dados respeita os intentos dos consumidores finais. Por exemplo, quando a oportunidade dos dados se torna crítica, pelas exigências de sistemas informáticos *real-time*, deverá ser prestada maior atenção à exactidão desses dados [Fischer & Kingma, 2001]. Por outras palavras, o melhor ou pior desempenho de cada uma das dimensões deve ser entendido num sentido mais abrangente e que vise o estabelecimento das exigências de qualidade dos dados advindas dos consumidores. Em [Bouzeghoub & Peralta, 2004] é perspectivado o relacionamento entre a frescura dos dados e as outras características dos dados. O relacionamento pode ser realizado segundo duas ópticas: a optimização da frescura dos dados em detrimento das outras dimensões ou o relaxamento da frescura dos dados em prol das restantes dimensões. A dificuldade coloca-se no modo de identificação das dimensões relacionadas e na concretização das técnicas, que prevejam o balanceamento adequado das dimensões, tendo em vista a melhoria do nível geral da qualidade dos dados.

### 2.4.6 Relacionamento das dimensões e os intervenientes nos dados

A associação entre as dimensões da qualidade dos dados e os diversos intervenientes nos dados, pretende salientar que o relacionamento entre o conhecimento por parte dos intervenientes e a qualidade dos dados difere consoante os papéis assumidos durante o processo de produção dos dados (produtores, administradores e consumidores). Os produtores fornecem a entrada inicial dos dados na organização. Os administradores são responsáveis pelo armazenamento e manutenção dos dados. Por fim, os consumidores utilizam os dados como suporte às suas actividades [Lee & Strong, 2004]. As investigações [Lee & Strong, 2003, 2004] identificam que o conhecimento sobre os processos de produção dos dados e a satisfação da qualidade dos dados, nas suas dife-

rentes dimensões, estão altamente correlacionados. Tendo em vista a obtenção de um maior nível de detalhe sobre este assunto, os estudos revelam igualmente três modos de conhecimento: *conhecer o quê*, *conhecer como* e *conhecer porquê*. Estes modos de conhecimento são cruzados com um conjunto de dimensões cruciais e representativas da qualidade dos dados: exactidão, oportunidade, completude, relevância e acessibilidade. Os estudos concluem que todos os modos de conhecimento interrelacionados com todos os processos de produção dos dados contribuem para a qualidade dos dados. Cada modo de conhecimento sobre os processos de produção dos dados, suportados por cada interveniente, contribui especificamente para as diversas dimensões da qualidade dos dados. É igualmente observável que o modo de conhecimento *conhecer porquê*, por parte do produtor dos dados, se trata do requisito mais importante para a alta qualidade dos dados nos processos de produção dos dados. Particularizando, o estudo apresenta fortes indicadores entre os intervenientes e as dimensões. Os produtores dos dados tendem a centrar-se na recolha dos dados relevantes, exactos e completos. Os administradores dos dados ocupam-se do armazenamento completo dos dados de modo a satisfazer as exigências ao nível da exactidão e oportunidade. Por fim, os consumidores dos dados pretendem identificar os dados suficientemente relevantes para a auxiliar na execução das tarefas em mãos. A figura 2-4 ilustra os resultados alcançados na investigação e salienta as diferenças sobre o conhecimento *conhecer porquê* pelos diversos intervenientes.

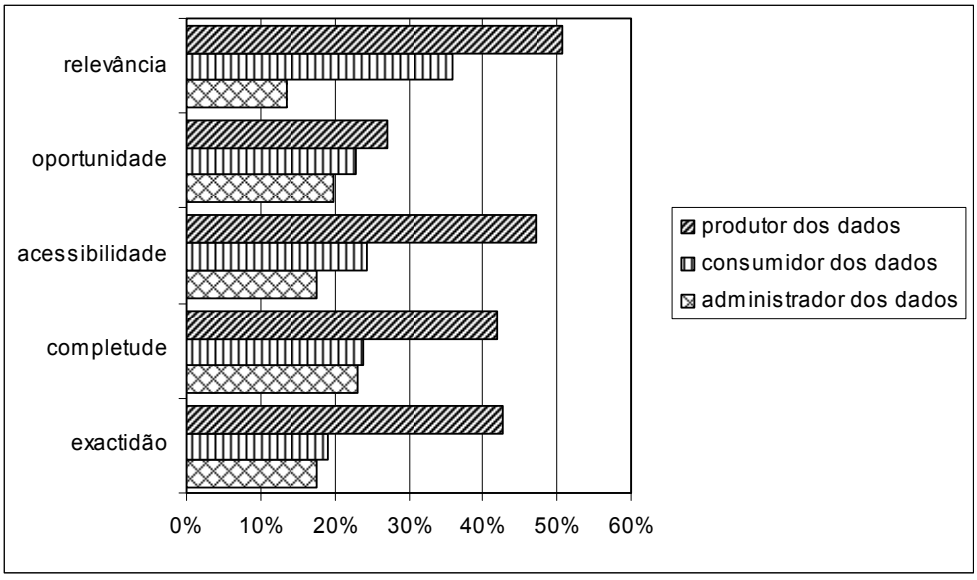



**Figura 2-4 – O conhecimento *conhecer porquê* cruzado com os intervenientes.**

## 2.5 O Impacto da qualidade dos dados

O impacto causado pela qualidade dos dados nas organizações pode ser observado segundo duas ópticas. A primeira procura reconhecer alguns benefícios facultados pela elevada qualidade dos dados nas organizações ou nos sistemas de informação que as representam. A segunda pretende apurar, ou pelo menos aferir, dos abalos sentidos pelas organizações que são resultado da diminuta qualidade dos dados. As consequências negativas provocadas pela fraca qualidade dos dados podem assumir um efeito directo na vida das organizações ou produzir um desfecho mais subtil, mas nem por isso menos desastroso na vida das mesmas. Assim, pretendemos arrogar ao sentido da importância vital que a qualidade dos dados assume nos sistemas de informação e por inerência, nas organizações [Strong et al., 1997].

### 2.5.1 A importância da qualidade dos dados

Os relatos de situações que revelam esta realidade são abundantes conforme faz notar a diversa literatura temática e proliferam sobre os mais variados assuntos e quadrantes da sociedade, quer no sentido mais amplo quer no sentido mais restritivo do termo. Os efeitos dos choques sentidos, negativos e positivos, têm-se mostrado mais intensos ao longo do tempo, extravasam o âmbito local, assumindo claramente, proporções globais e fazem-se sentir, indiferentemente, em organizações privadas e públicas. Outrora, os danos e as oportunidades surgidas eram explicadas por outros factores que não a qualidade dos dados, como sejam entre outros, a implementação de novas tecnologias, os reajustes organizacionais, as mudanças culturais, a aplicação de novas metodologias e de filosofias de gestão. A aceitação da informação como recurso estratégico e fonte criadora de vantagens competitivas, imprimiu uma alteração no modo como os dados passaram a ser encarados. Estes começaram a ser entendidos como o único património inalienável e vital para as organizações. A propensão para a focalização no tratamento dos dados deve-se à circunstância de estes serem percepcionados não como uma consequência das actividades das organizações, mas antes pela consciencialização da sua importância. Esta importância é, especialmente, motivada pelo facto das organizações estarem mais dependentes dos seus sistemas de informação e por se encontrarem mais expostas perante um meio envolvente cada vez mais atento e punitivo. Em [Eckerson, 2002], esta tendência é explicada pelo deslocamento da focalização na economia industrial para a economia da informação. A economia da informação consiste num cenário em que as organizações competem pela capacidade em receber e emitir informações e não apenas na produção de bens. O mesmo estudo refere um gestor duma organização, “A nossa companhia vende dados. Se não estiverem correctos, não existiremos”.

Os dados conduzem todas as operações, assumindo-se como o *input* fulcral na tomada de decisão [Redman, 2004]. A quantidade de dados adquiridos e armazenados pelas organizações tem crescido a elevado ritmo. Em [Redman, 2004] é citado um colaborador da IBM: "na IBM temos 10 vezes mais contactos com pessoas, 100 vezes mais velocidade na rede, 1000 vezes mais dispositivos tecnológicos e um milhão de vezes mais dados". Adicionalmente, é salientado para o maior número de dados publicados na Internet e não é vislumbrado o fim desta tendência. O recurso a maiores quantidades de dados pode provocar um maior volume de falhas na sua qualidade. Em [Lee et al., 2000b] considera-se que o exponencial incremento na utilização da Internet, facilita o processamento e a distribuição rápida e descontrolada da fraca qualidade dos dados.

Hoje em dia, caminha-se para a aceitação dos dados e da sua qualidade como a origem de grandes males ou de possíveis sucessos. Os resultados obtidos, anteriormente imputados a outros factores, têm sido agora manifestamente focalizados em torno da qualidade dos dados. A tomada de decisão assente apenas nas informações disponíveis, muitas delas de duvidosa correcção, ausentes de consistência, oportunidade desadequada e insuficientemente relevante para a responsabilidade inerente ao processo de decisão, conduz a fortes embates, provavelmente, de terríveis consequências na vida das organizações. Esta constatação revela-se ainda mais dramática, dada a inconsciência destas questões pelos responsáveis organizacionais [Redman, 2004]. Assim, é urgente que as diversas entidades repensem estes assuntos, em especial, aquelas que funcionam como agentes reguladores. A opção por posições mais radicais que visem, além dos reforços orçamentais, a iniciativa de medidas legislativas que reforcem a orientação neste domínio, é algo recentemente desenvolvido conforme iremos referir mais adiante neste capítulo.

### 2.5.2 Macro impacto

A preocupação com a qualidade dos dados resulta de constantes investigações e estudos que desnudam as verdadeiras razões de muitos dos sucessos ou insucessos das organizações e admitem que a qualidade dos dados é dos problemas mais críticos que as organizações enfrentam nos nossos dias. Estas investigações explicam a origem de consequências trágicas por razões oriundas da deficiente qualidade dos dados. Muitas destas consequências derivam de processos de tomada de decisão erráticos com graves reflexos no âmbito político, cultural, social e económico, tanto numa perspectiva local como global.

A investigação [Fischer & Kingma, 2001], sobre os desastres ocorridos com o *space shuttle Challenger* e o avião de passageiros iraniano abatido pelo navio norte-americano *Vincennes*, revela que estes trágicos acontecimentos não ficaram a dever-se, exclusivamente, aos motivos oficiais

amplamente invocados e é apontada a existência de factores verificáveis e relacionados com a qualidade dos dados. A investigação mostra, igualmente, que em ambos os casos a garantia de uma boa qualidade dos dados não figurava no topo das prioridades das referidas entidades. Os dois sistemas de informação apresentavam dificuldades em dimensões cruciais dos dados, sendo identificados cerca de 10 problemas em seis dimensões na *National Aeronautics and Space Administration* (NASA) e 8 problemas em cinco dimensões na marinha norte-americana. O estudo [Wang et al., 2003], sobre o ataque terrorista às torres gémeas de Nova Iorque, em 11 de Setembro de 2001, observa que, para além da identificação dos problemas mais comuns, a existência de problemas relacionados com a qualidade dos dados era uma realidade. Em especial, a falha de informação vital facultada para a tomada das decisões certas por parte dos responsáveis. Neste sentido, o estudo, inclusivamente, aponta uma aproximação, *Total Information Awareness*, capaz de assegurar um alto nível de qualidade dos dados no combate ao terrorismo. Ainda sobre o mesmo assunto, idêntica opinião é partilhada em [Redman, 2004] e recomenda a necessidade em possuir dados correctos, oportunos e disponíveis nos locais próprios. Após o atentado, descobriu-se que durante o ano e meio que precedeu o ataque, foram criadas pelos terroristas, usando falsos nomes, 35 contas em bancos [Hudicka, 2002]. Em [Redman, 2004], são identificados outros casos igualmente inquietantes e que se devem a questões de natureza dos dados. Entre estes, a eleição do presidente *Bush* no ano 2000, que resultou de um complexo processo de contagens e recontagens dos boletins de voto, denotando nítidas quebras de certeza dos dados e o ataque à embaixada chinesa durante a guerra da Bósnia. Sobre o ataque à embaixada chinesa, um outro estudo [Hussain & Beg, 2003] revela que os dados possuídos pela *Central Intelligence Agency* (CIA) sobre o alvo definido se encontravam desactualizados. O mesmo estudo apresenta um caso no domínio da saúde que aponta a fraca qualidade dos dados como a causa da morte de cerca de 98000 pessoas anualmente nos E.U.A. (e.g. através da incorrecta prescrição de medicamentos). Ainda o mesmo estudo, refere a dificuldade dos serviços de inteligência norte-americanos em compreenderem o significado e as ligações dos nomes árabes. A recolha independente de nomes e nomes alternativos para as bases de dados, efectuada por cada um dos serviços de inteligência (*Federal Bureau of Investigation* (FBI), CIA, *Immigration and Naturalization Service* (INS) e outras agências) gera ainda mais confusão sobre o assunto. A CIA admite existirem cerca de 60 nomes alternativos para o líder da Líbia *Moammar Gadhfi* (figura 2-5). As dificuldades sentidas na identificação concreta de uma personalidade sobejamente conhecida, presidente de um país e possuidor de um passado associado a ligações terroristas, são apenas a "ponta do iceberg" no domínio da qualidade dos dados. Estas questões explicam as recentes medidas tomadas pelo presidente dos E.U.A., que apontam a criação de um serviço de informação directamente dependente da casa-branca e coordenador de todas as agências e serviços de inteligência. Este novo serviço deve ser

capaz da conciliação e integração de dados vitais para administração *Bush* e irá traduzir-se num importante passo para a opção estratégica de segurança nacional antiterrorista.

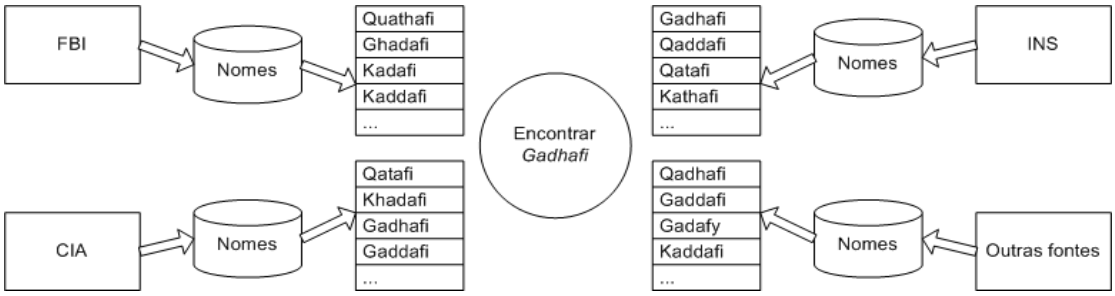



**Figura 2-5 – Dados nas agências governamentais [Hudicka, 2002].**

### 2.5.3 Micro impacto

Outros exemplos, como os acima apresentados, começam a abundar na literatura sobre qualidade dos dados. Geralmente, os estudos expõem situações mais simples e de domínio mais restritivo, no domínio organizacional ou tecnológico. No estudo [Eckerson, 2002] é referido que 40% das organizações sofreram problemas e custos devido à fraca qualidade dos dados. O estudo relata, igualmente, alguns casos ocorridos em diversas organizações que se traduziram no desaproveitamento de oportunidades ou na implicação de elevados custos para as mesmas:

- Um banco detectou que cerca de 2/3 dos empréstimos à habitação concedidos estavam incorrectamente calculados.
- Uma seguradora esperou dois anos pela implementação do seu sistema de suporte à decisão devido a problemas detectados com a qualidade dos dados.
- Uma firma de telecomunicações perdeu cerca de 7 milhões de euros num mês por causa de dados incorrectamente introduzidos.

Em [Olson, 2003], mostra-se a falta de qualidade dos dados como a causa do aumento dos custos ou do desperdício de oportunidades em inúmeras áreas: os custos de refazer os trabalhos (e.g. o tempo extra necessário para reconciliar os dados); o acréscimo de custos na implementação de novos sistemas (e.g. a implementação de SDW demorar mais que o previsto no plano); os atrasos na entrega de informações para a tomada de decisão (e.g. sem informações sobre o mercado concorrencial, os estrategas de marketing têm dificuldades na execução de políticas adequadas e oportunas); a perda de clientes devido ao fraco serviço prestado (e.g. a insatisfação dos clientes e os problemas na cadeia de suporte à produção). Adicionalmente, em [Eckerson, 2002] é apontada a perda de credibilidade no sistema informático (figura 2-6). Outro estudo [Redman, 2004] refere

os danos associados à degradação da imagem da organização. Em [English, 1999] é listada uma série de problemas quotidianos nas organizações decorrentes da fraca qualidade dos dados. O estudo promovido por uma consultora [Kenyon et al., 2000], realizado em 600 organizações, corrobora os estudos apresentados. Saliente-se o facto de 75% das organizações denotar problemas substanciais resultantes de dados defeituosos, de metade das organizações apresentarem custos extraordinários resultantes da fraca qualidade dos dados e cerca de um terço demorar a implementação de novos sistemas informáticos.

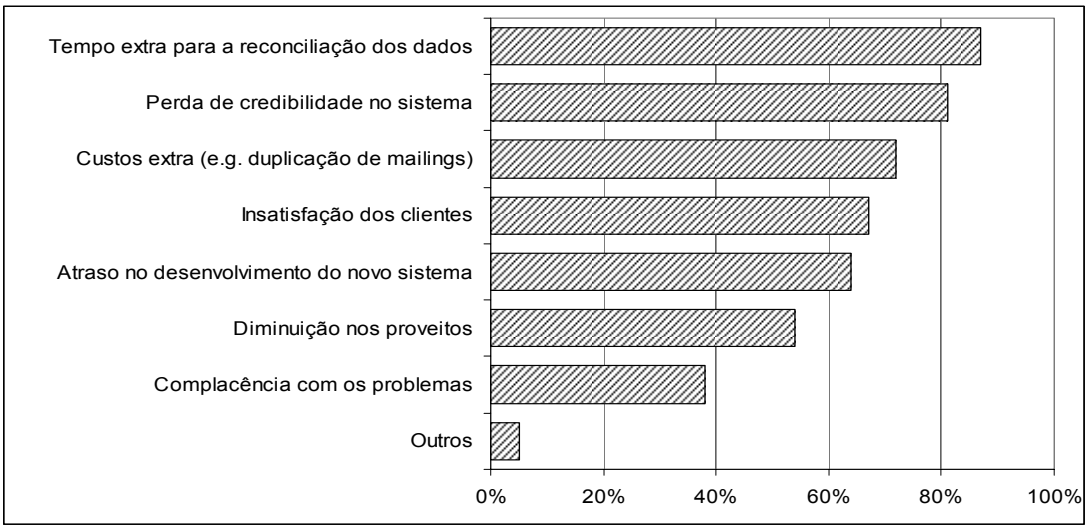



**Figura 2-6 – Problemas causados pela fraca qualidade dos dados [Eckerson, 2002].**

Noutra perspectiva, em [Redman, 1998] são visualizados os impactos comuns provocados pela fraca qualidade dos dados nas diversas camadas da estrutura organizacional: plano operacional, plano táctico ou de coordenação e plano estratégico. No plano operacional, algumas das consequências provocadas pela fraca qualidade dos dados são: a pouca satisfação dos colaboradores, a pouca satisfação dos clientes e o aumento dos custos. No que respeita ao embate provocado no plano táctico ou de coordenação detectam-se as fracas e demoradas tomadas de decisão, a maior dificuldade em implementar sistemas de suporte à decisão, um aumento da desconfiança organizacional e uma maior dificuldade na reengenharia de processos organizacionais. Por último, o impacto ao nível estratégico revela-se pela maior dificuldade na definição e execução das estratégias, pelo desvio da atenção dos gestores e consequente comprometimento da possibilidade de alinhamento com as organizações, bem como, em contribuir para questões de natureza política sobre os dados. Ainda no plano estratégico, em [Eckerson, 2002] é considerado o modo como a fraca qualidade dos dados mina os planos ou projectos estratégicos porque estes se baseiam em dados referenciais imprecisos ou incorrectos.

Tentando-nos reportar a um plano estritamente técnico, mas com graves repercussões para a vida das organizações, podemos observar em diferentes investigações a quantidade significativa de dados incorrectos ou impróprios, que são continuamente armazenados nas bases de dados das organizações. Estes dados (bons e maus) uma vez processados e acedidos corroboram a expressão, amplamente referenciada, *garbage in, garbage out*, ou seja, os dados defeituosos armazenados nos sistemas informáticos conduzem a decisões erráticas. O estudo [Redman, 1998] aponta para erros na ordem dos 5% nas bases de dados das organizações. A existência de dados incorrectos nas bases de dados pode constituir-se como um obstáculo para as organizações. Todavia, interessa salientar que esses defeitos não podem ser integralmente retirados, situação praticamente impossível e desnecessária de realizar [Ballou et al., 2004]. Impossível porque mesmo supondo que se possui dados 100% correctos, essa garantia não é contínua e duradoura, ou seja, os dados deterioram-se por si mesmos instantaneamente (e.g. o cliente muda de habitação ou de estado civil). É igualmente, uma situação desnecessária, porque os esforços associados (meios materiais e humanos) na tentativa de perfeição dos dados podem tornar-se injustificáveis, tanto em termos financeiros como temporais. Os dados devem situar-se num plano de qualidade que sirva as exigências presentes e futuras das organizações. Em suma, as organizações encontram-se sujeitas a constantes evoluções, originando mudanças nas necessidades e desejos sentidos em termos de informação, intimando estas a serem correctas, correntes, integradas e adaptáveis na resposta adequada às contingências das actividades de negócio [Brackett, 1996].

### 2.5.4 Os custos da fraca qualidade dos dados

O impacto gerado pela fraca qualidade dos dados toma proporções mais ou menos graves na vida das organizações. Sendo que essas consequências, geralmente, são proporcionais à dimensão ou categoria da organização em causa. Ao pensarmos no governo dum país como uma organização em sentido lato, verificamos que as sequelas provocadas pelos defeitos ou ausências dos dados podem assumir amplitudes trágicas. Os prejuízos daí resultantes, medidos em termos de unidades monetárias, podem situar-se em verbas admiravelmente incríveis. Por exemplo, os E.U.A. compensaram em mais de 22 milhões de euros o governo chinês pelo ataque acidental sobre a sua embaixada na Bósnia. Porém, associado a este montante, outros danos mais sérios são impossíveis de mensurar, como sejam a imagem dos serviços de inteligência e das unidades militares envolvidas, em sentido restrito, e a própria imagem do governo norte-americano, em sentido lato [Redman, 2004]. A câmara de comércio norte-americana [3] estima custos anuais imputados ao estado norte-americano na ordem dos 700 biliões de euros, em resultado de manifestações de fraca qualidade dos dados. Esta estimativa resulta dos aumentos dos bens e serviços, dos incre-

mentos das taxas, dos baixos salários, do maior desemprego, da estagnação económica e da pouca inovação tecnológica.

Reportando a organizações mais pequenas, os efeitos apesar de menos intensos são semelhantes. *English* alerta para a necessidade da determinação dos custos envolvidos originados pela fraca qualidade dos dados, para isso, recorre a um processo da *Total Information Quality Management* (TIQM) capaz de efectuar a contabilização dos custos. Salienta, igualmente, a importância da quantificação dos custos, na medida em que possibilita: a determinação dos reais impactos provocados pela fraca qualidade dos dados no negócio; permite tomar iniciativas que visem a melhoria da qualidade dos dados e serve de base para avaliar as iniciativas de melhoria da qualidade dos dados [English, 1999] [English, 2003a]. A efectiva avaliação dos custos exige o entendimento sobre as formas de incorrecção dos dados e os custos resultantes, sendo que, basicamente, a maioria dos problemas derivam da perda de desempenho de uma ou mais dimensões basilares dos dados [Cappiello & Francalanci, 2002]. A falha ao nível da dimensão exactidão poderá gerar a perda de confiança interna ou a comunicação de informações imprecisas (e.g. quartos de hotel vendidos ao preço de 59 euros em vez do real valor de 259 euros ou bilhetes de avião vendidos a 5 euros) [Redman, 2004]. A ausência de dados pode afectar a dimensão completude, provocando o desperdício de oportunidades ou a tomada de más decisões. Em relação às restantes dimensões podem ser observados comportamentos desajustados em relação às necessidades de informação e que suscitam certamente prejuízos, mais ou menos quantificáveis, nas organizações.

A obtenção dos custos reais é de difícil determinação, porque não é possível cingirmo-nos meramente aos custos tangíveis ou contabilísticos. Em geral, as verbas divulgadas consideram os custos tangíveis, mais praticáveis de calcular e que resultam do esforço de detecção, da resolução dos erros e da duplicação das actividades [Redman, 2004]. Inerentemente aos prejuízos determináveis, uma outra classe de custos, mais ténue, mas significativamente mais violenta é passível de ser observada, são os designados custos intangíveis, também designados como incalculáveis ou ocultos. Estes são praticamente indetermináveis ou mesmo impossíveis de obter e são, por vezes, difíceis de ser verificados. Esta classe de custos respeita, nomeadamente, a questões de natureza sócio-organizacional e psico-organizacional como sejam: a moral dos funcionários, a desconfiança interna, o desaproveitamento de oportunidades, a motivação organizacional, as dificuldades de alinhamento com a estratégia da organização, a responsabilização das acções tomadas e a imagem da organização na sociedade.

Comummente, na literatura temática verifica-se a adopção de um valor percentual, indicativo do montante equivalente aos danos causados às organizações pela utilização de dados de fraca qua-

lidade. Em [Redman, 1998], os custos tendem a representar entre os 8% e os 12% dos ganhos, podendo em certos casos atingir valores entre os 40% a 60%. Os custos associados à má qualidade dos dados, segundo *English*, representam entre os 10% e os 25% dos proveitos ou do orçamento total da organização (e.g. custos irrecuperáveis, custos de refazer tarefas, produtos ou serviços) [English, 1999]. Em [Olson, 2003] estima-se os custos provocados pela fraca qualidade dos dados no intervalo entre os 15% e os 25% dos proveitos operacionais. Mais recentemente, em [Redman, 2004], é estabelecido que apesar do valor referencial respeitar os 10% dos proveitos, estudos recentes apontam que um valor em torno dos 20% é uma estimativa mais razoável na realidade. O mesmo estudo descreve, igualmente, a regra 1 para 10, isto é, se os dados estão bons então o custo duma operação é de uma unidade monetária, se a qualidade dos dados é pobre, então o custo da operação será de dez unidades monetárias. Abstraindo-nos da volatilidade dos valores percentuais apresentados pelos diferentes estudos, parece sintomático, que os custos provocados pela fraca qualidade dos dados assumem uma fatia significativa dos proveitos organizacionais. Acresce ainda, que em [Redman, 2004] os custos ocultos representam uma fatia de montante semelhante aos custos tangíveis. Em suma, podemos constatar que o conjunto destes valores supera várias vezes os investimentos comuns necessários para a melhoria da qualidade dos dados. Perante este panorama, a opção é de sentido único e consiste no esforço contínuo em busca da obtenção da melhor qualidade dos dados possível.

### 2.5.5 Os benefícios da qualidade dos dados

A determinação dos custos mostra-se uma tarefa de difícil quantificação, semelhante ou pior cenário, é notado na obtenção dos benefícios resultantes da melhoria da qualidade dos dados, porquanto compreendem não apenas os benefícios tangíveis, como aqueles de mais difícil avaliação, os intangíveis. Conforme referido em [Watson et al., 2001], a análise dos investimentos realizados é geralmente uma tarefa problemática, porque exige a atribuição de valores aos benefícios. O estudo da consultora *PriceWaterhouseCoopers* [Kenyon et al., 2000] constata que enquanto os custos são específicos, os benefícios tendencialmente são mais genéricos. No estudo cerca de 3/4 das organizações consideraram que os investimentos ao nível dos dados trouxeram um impacto positivo às organizações, ronda os 60% as organizações que manifestaram corte nos custos de processamento e mais de 40% viram impulsionadas as suas vendas devido à realização de melhores análises sobre os dados dos seus clientes.

O investimento na melhoria da qualidade dos dados traduz-se na obtenção de benefícios tangíveis e intangíveis. Normalmente, os benefícios alcançados devem-se à resolução dos problemas e custos resultantes da fraca qualidade dos dados, como sejam o aumento da satisfação dos clien-

tes, a existência duma única versão dos dados, a elevada confiança nos sistemas informáticos, a redução de custos, entre outros [Eckerson, 2002]. O mesmo estudo recorre ao exemplo duma instituição financeira que, pela introdução de um projecto de melhoria da qualidade dos dados, obteve poupanças de cerca de 100 mil euros anualmente. Este projecto resultou num *Retorno do Investimento* (ROI) de 188%. Sobre os benefícios proporcionados pela melhoria da qualidade dos dados, um estudo [English, 2003a] apresenta o caso de sete organizações que, no seu conjunto, pouparam cerca de 400 milhões de euros. Este valor resulta de economias ao nível dos diversos recursos materiais, humanos e outros custos directos exigidos pela falha dos processos de tratamento dos dados. Sendo que os benefícios intangíveis, em alguns casos, superaram os benefícios directos. Os investimentos necessários para debelar as fraquezas da qualidade dos dados representam uma ínfima parte dos potenciais benefícios, logo parece plausível a obtenção de elevados índices de retorno, como se comprova pelo caso duma organização que obteve um ROI na ordem dos 700% [English, 2003a]. Em [Kenyon et al., 2004] é referido que o retorno da introdução de iniciativas de melhoramento dos dados ronda 10 vezes o investimento realizado.

## 2.6 Razões da fraca qualidade dos dados

Os problemas de qualidade dos dados verificam-se em organizações quer de âmbito público quer de âmbito privado. Nenhuma organização se encontra imune a este tipo de ocorrências [Hudicka, 2002]. Os motivos causadores desta problemática têm sido perspectivados de modo complementar por diversas investigações. Segundo *Brackett* as origens dos problemas da qualidade dos dados podem ser agrupadas em quatro categorias [Brackett, 1996]. A primeira categoria apela para a falta de consciência dos diversos intervenientes sobre os dados organizacionais. A segunda categoria diz respeito à compreensão dos dados existentes, ou seja, mesmo que os dados estejam documentados e sejam conhecidos, a noção sobre o seu real conteúdo e significado fica, geralmente, por compreender. A terceira classe da origem de problemas com a qualidade dos dados, consiste nos designados dados disparatados[4], ou seja, a frequente variabilidade do conteúdo, do significado e do formato nos valores referentes aos mesmos dados, inviabiliza a utilização desses dados [Brackett, 1996]. A acumulação, ao longo dos anos, destes dados disparatados origina o dominado caos dos dados. A última categoria corresponde à existência de redundância nos dados e pretende realçar a dificuldade na determinação do local que apresenta os dados mais correntes e que reflectem melhor o mundo real. Em [Olson, 2003], são apontados problemas de ordem operacional no manuseamento dos dados e enuncia quatro ordens de factores: a entrada incorrecta

[4] Tradução do inglês: *disparate data*

dos dados, a decadência dos dados, o transporte e reestruturação dos dados e a utilização dos dados. A entrada incorrecta dos dados pode ficar a dever-se a erros de introdução dos dados (e.g. o operador digita 15€ em vez de 51€), a falhas nos processos de entrada de dados (e.g. um formulário que contenha um campo que não especifica se deve ser inserida a freguesia de nascimento ou de residência) e a erros deliberados (e.g. o desconhecimento da informação real sobre o rendimento *per capita* do agregado familiar dum aluno, pode gerar o preenchimento voluntário de dados incorrectos). No que respeita à decadência dos dados, é possível constatar a necessidade de reverificações dos dados em intervalos mais ou menos regulares (figura 2-7). Pois, se num dado instante, um conjunto de dados pode-se apresentar correcto, num ápice revela-se a existência de deteriorações desses dados e que se explicam por alterações normais dos valores dos dados (e.g. a alteração do estado civil de um cliente). Um terceiro factor causador de originar problemas com os dados resulta do transporte e reestruturação dos dados entre sistemas informáticos, como seja os dados em trânsito no processo de *Extracção e Transformação e Carregamento dos dados* (ETL), provenientes do SO, podem sofrer acções de tratamento incorrectas.

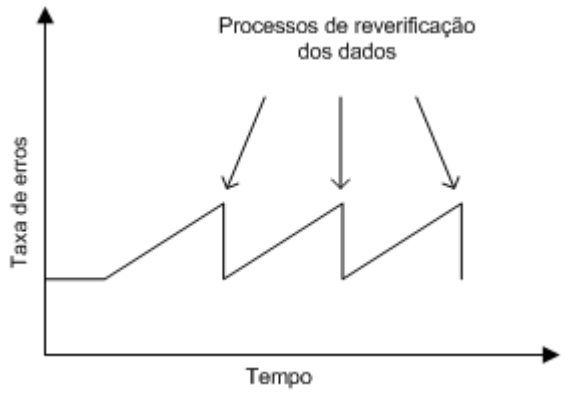



**Figura 2-7 – A decadência dos dados correctos [Olson, 2003].**

O estudo do *The Data Warehouse Institute* [Hudicka, 2002] vai um pouco mais além e justifica a opção consciente por parte das organizações em possuírem problemas ao nível da qualidade dos dados. Conforme visto anteriormente, o FBI, o INS e a CIA possuem métodos distintos que devolvem valores diferentes sobre o mesmo nome (cf. Figura 2-5). Em primeiro lugar, os custos incomportáveis na substituição dos sistemas informáticos existentes, tanto em termos financeiros directos, como indirectos (e.g. o sistema informático existente está enraizado pelos funcionários). Um segundo motivo considera a possibilidade de construção de um interface sobre a arquitectura existente, mas esta apresenta dificuldades de nível logístico e de entendimento dos sistemas. Estas razões levam a uma terceira questão impeditiva que consiste na incapacidade em gerar um ROI apelativo. Em [Ballou & Tayi, 1998] consideram-se que algumas razões originárias da fraca quali-

dade dos dados surgem pela combinação de dados que não foram projectados para serem integrados; pela baixa prioridade em assegurar a qualidade dos dados, apesar de reconhecida a sua importância e a difícil determinação da natureza das deficiências dos dados. A pesquisa elaborada pelo *Aberdeen Group*, registou as causas típicas que provocam a quebra de qualidade na informação [Kimball et al., 1998]. É observável que o factor humano (erros introduzidos e as omissões de valores dos dados) constitui-se como a principal razão originária da perda dos dados, ou seja, a qualidade das suas acções é reflectida directamente no impacto infringido às organizações (figura 2-8). Esta constatação é o factor mais influente na qualidade dos dados e por isso, especial atenção deve ser orientada no desenvolvimento da motivação e preparação adequadas dos interlocutores com os sistemas informáticos.

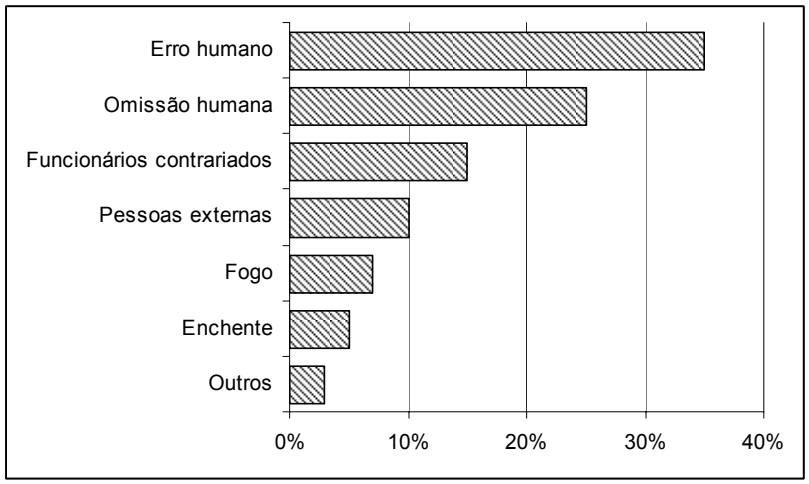



**Figura 2-8 – Causas da perda dos dados [Kimball et al., 1998].**

## 2.7 Tendências da qualidade dos dados

As organizações começam a ficar atentas quanto às problemáticas relacionadas com os dados. Algumas investigações fazem notar a necessidade prioritária em possuir dados de elevada qualidade. O estudo [Wang, 2004], sobre o estado de alerta das organizações para os problemas que envolvem os dados, refere que estas são, actualmente, cada vez mais conhecedoras das reais vantagens advindas da boa qualidade dos dados, já que têm a experiência do sofrimento, por elevadas perdas, devido à diminuta qualidade dos dados. Na verdade, as organizações deixam de situar-se num estado de profunda letargia ou incompreensível inconsciência sobre os problemas relativos à qualidade dos dados, para assentarem num novo patamar, consciente e em alerta para os problemas, mas mantendo-se incapazes na sua resolução. Este desassossego é possível ser verificado num estudo [Paulson, 2000] revelador das intenções prioritárias das organizações em

questões relacionadas com a implementação de infra-estruturas tecnológicas. Nesse estudo observa-se que a melhoria da qualidade dos dados relativas aos seus clientes está no topo da lista das prioridades. Este é sem dúvida um passo significativo, porque revela a tomada de consciência pelos gestores das organizações para o papel de destaque assumido pelos dados no eventual sucesso dos sistemas de informáticos e em consequência das organizações. A investigação [Scannapieco & Catarci, 2002] observa que o nível de qualidade é importante tendo em vista o fornecimento de bons serviços tanto do domínio público como do domínio privado. Para isso, recorre ao caso em que a recente legislação italiana prevê a possibilidade dos italianos, radicados em países estrangeiros, votarem nas eleições. Esta medida constitui-se como a base justificativa para a tomada de iniciativas relacionadas com a qualidade dos dados. Para a efectivação da concretização deste objectivo houve necessidade de actualização da morada dos cidadãos votantes.

A disposição, por parte das organizações, em centralizar os devidos esforços para garantir a melhor qualidade dos dados possível, tem sido correspondida pelas entidades competentes, quer governamentais quer de natureza associativa. Ou noutra perspectiva, as entidades competentes têm tentado regular este domínio, através de legislação sobre a qualidade dos dados e tomando a iniciativa em a fazer adoptar pelos seus serviços e em consequência pelas organizações envolventes, visando assim atingir os desejos dos consumidores de informação. Estas iniciativas são classificadas em [English, 2003c] como formas de garantir a "pureza" da informação. Uma dessas iniciativas foi desenvolvida pelo sector privado financeiro visando a resposta às preocupações dos seus clientes na forma como os seus dados pessoais eram tratados pela classe de instituições em causa. Esta iniciativa designada *Gramm-Leach-Bliley* propôs regular a partilha de informações pessoais relativas aos clientes entre as instituições filiadas na classe e entre estas e outras organizações não filiadas [English, 2003c].

Outra iniciativa, de âmbito público, conhecida por *Data Quality Act*[5] foi aprovada pelo congresso norte-americano em Dezembro de 2000. Esta iniciativa é motivada pelos elevados custos causados pela pobre qualidade dos dados às organizações públicas, às organizações privadas e aos cidadãos em geral e pretende que as queixas dos consumidores relativas à qualidade dos dados forcem a tomada de acções legais visando o seu melhoramento [3] [English, 2003c]. Trata-se dum mecanismo para assegurar que as decisões tomadas pelo governo sejam baseadas em informações de alta qualidade [Kyl, 2005]. Neste sentido, foi previsto que a OMB, emitisse um despacho sobre as directrizes para as agências federais norte-americanas, de modo que estas assegurem a

[5] *Data Quality Act Public Law 106-554, H. R. 5658, Section 515 requires Federal*

qualidade, objectividade, utilidade e integridade dos dados propagados ao público [Copeland & Simpson, 2004]. A medida previa que, no terceiro trimestre de 2002, as agências federais estabelecessem os seus próprios procedimentos em vista garantir as directrizes gerais para assegurar a disseminação da informação e permitisse aos cidadãos o confronto sobre dados que estes considerem não respeitar essas directrizes. Podendo a decisão final ser tomada pelo tribunal em caso de se verificar conflito [4]. Sobre a iniciativa, o *New York Times*, considera que o governo mais poderoso do mundo tem de emitir informação de qualidade. Para o vice-presidente da Câmara de Comércio norte-americana, esta medida corresponde à maior iniciativa em termos regulamentais e terá um impacto tão intenso que é difícil imaginar o seu alcance [3] [5]. Assim, as organizações passam a ter necessidade em fornecer a qualidade dos dados esperada aos consumidores ou correm o risco da legislação forçar a garantia dessa qualidade [English, 2003c]. A iniciativa promovida pela OMB salienta a necessidade da elevada importância da qualidade dos dados nas organizações, bem como a forma de encarar os meios precisos para que concretização desse vital objectivo seja uma realidade. Para isso, dois princípios parecem subjacentes: o envolvimento de todos os intervenientes das organizações e encarar o problema da qualidade dos dados como um processo infindável e em contínuo desenvolvimento [English, 2004]. A manutenção e melhoramento da qualidade dos dados é um processo que nunca termina. Inerente a estes dois princípios, podemos encontrar preocupações de âmbito transdisciplinar na gestão da vida das organizações, como sejam: promover a mudança cultural, incutir o sentido de responsabilização dos colaboradores, implementar novas filosofias de gestão, executar novos métodos de gestão dos recursos humanos e reestruturar procedimentos e políticas.

Do ponto de vista tecnológico, as questões relativas à qualidade dos dados têm merecido, igualmente, bastante atenção, conforme faz notar um recente estudo [Naumann & Roth, 2004]. Nesse estudo, é-nos mostrada a evolução dos *Sistemas de Gestão de Base de Dados* (SGBDs), em especial, no diz respeito à garantia da qualidade dos dados e à amabilidade para os utilizadores, ou seja, devem assegurar a alta qualidade no armazenamento dos dados e facilitar o acesso aos mesmos. É verificada a crescente preocupação na introdução de ferramentas que validem, armazenem, manipulem, limpem e devolvam os dados de uma maneira facilmente interpretável. O estudo examina os SGBDs relativamente ao suporte da qualidade dos dados e para isso, recorre a um conjunto de critérios de qualidade dos dados que os SGBDs devem respeitar. Estes critérios correspondem às diferentes dimensões associadas aos dados e que devem ser o mais eficaz e eficientemente garantidas. Ao princípio *garbage in garbage out* associado aos sistemas que processam dados, o estudo pretende evoluir no sentido de postular o princípio *quality in quality out* para os modernos e bem construídos SGBDs.

# Capítulo 3

# Qualidade dos dados em SDWs

O DW é definido, clara e objectivamente em [Inmon, 1996], como um sistema de dados orientado por assuntos, integrados, não voláteis e variantes no tempo. Estas características quando devidamente exploradas, afirmam o DW como um meio tecnológico capaz de conceder vantagens estratégicas às organizações no que respeita à tomada de decisão, por isso, interessa que a sua implementação seja um sucesso e a sua manutenção uma realidade. O sucesso destes sistemas de dados prevê a materialização de um conjunto de pressupostos, como sejam entre outros: o alinhamento com a estratégia da organização; a engenharia de requisitos adequada; a solidez e disponibilidade tecnológica que alicerça o DW; as técnicas de modelação utilizadas e a sustentação em processos de extracção e refrescamento dos dados de modo eficaz e eficiente. A garantia duma elevada qualidade dos dados, nas mais variadas dimensões, desde as fontes de dados até à disponibilização dos dados trabalhados aos utilizadores finais, assume-se igualmente, como factor contributivo para o sucesso destes sistemas.

Os SDWs compõem-se de recursos e meios tecnologicamente integrados em vista a disponibilização ao consumidor final de informação preciosa que o auxilie e o proteja no exercício da tomada de decisão. Por consumidor final da informação, pretende-se referenciar todos aqueles que beneficiam de tal sistema como suporte para as suas actividades e funções dentro das organizações. Atendendo às características dos SDWs, os principais utilizadores destes sistemas são aqueles que arcam com as maiores responsabilidades dentro duma organização. Neste sentido, as informações fornecidas aos consumidores finais devam roçar a senda da excelência. Caso contrário, podem significar a inutilidade de tal sistema ou o descalabro da própria organização quando confiante em informações erradas produzidas e apresentadas pelo sistema.. Assim, a informação

produzida e resultante deve tender para a excelência. Por excelência não pretendemos considerar a informação como um patamar estático a atingir, mas antes um grau de qualidade dos dados, que as organizações devem continuamente perseguir, em vista disponibilizar sempre uma qualidade dos dados que satisfaça ou supere as expectativas dos utilizadores finais, ou seja, capaz de produzir a diferença pela sua utilidade na tomada de decisão.

Ora, as vantagens resultantes da utilização dos SDWs conduzem a que estes possam ser assumidos como a pedra angular da estratégia e suporte de tomada de decisão das organizações. É por isso, imperativo que os dados constantes no DW sejam de facto uma mais valia benéfica para as organizações. Assim, podemos depreender que um SDW é tão bom quanto os dados nele contidos [18]. A permanência de dados de má qualidade no sistema implica que o consumidor desconfie das informações e análises divulgadas por este. Neste sentido, aos dados e aos processos que os manipulam deve ser arrogado o papel central de entre os diversos recursos e meios envolvidos num SDW. A obtenção de ROI num SDWs pode ser analisada pela utilidade e elevada qualidade dos dados disponibilizados. Daí que seja compreensível o elevado número de insucessos dos SDWs quando tal não acontece. A história recente destes sistemas mostra que a hipótese de singrarem na sua plenitude é manifestamente reduzida [Watson et al., 2001]. Os prejuízos originados pela fraca qualidade do sistema, nomeadamente aqueles respeitantes à qualidade dos dados, apresentam-se geralmente elevados e muitas vezes incomportáveis. Segundo *Brackett*, os maiores problemas na implementação dos SDWs e no fornecimento de um suporte à decisão adequado aos decisores prendem-se com questões relativas aos dados, como sejam: as discrepâncias entre os dados existentes, as dificuldades em desenvolver e gerir níveis de agregação adequados e a incapacidade em fornecer um suporte à decisão compreensível [Brackett, 1996].

## 3.1 O SDW como sistema de suporte à decisão

Desde sempre que o processo de tomada de decisão se apresenta um exercício corrente nos diversos patamares da estrutura orgânica das organizações. Todavia, as dirigidas pelos principais gestores estão incumbidas, naturalmente, de um maior grau de responsabilidade, pois se o seu sucesso faz corresponder a concretização de proveitos ou benefícios para as organizações, o insucesso da prática decisória implica avultadas perdas ou mesmo o desmoronamento da organização. Actualmente, a globalização económica, cultural e social impulsiona as organizações a maiores exigências ao nível da qualidade das decisões tomadas num mercado evoluindo em todas as suas vertentes. A abolição fronteiriça associada a grandes incrementos em meios tecnológicos não concede segundas oportunidades num teatro de operações concorrencial, autoritário e em constante mutação. A primeira função dos gestores consiste em tomar decisões e estas podem

ser separadas em três categorias distintas: estratégicas, de coordenação e operacionais. Enquanto que as duas últimas categorias se reflectem a nível interno e restrito na organização, a primeira categoria consiste na interacção da organização com o meio envolvente (o mercado e a sociedade) [Rascão, 2000]. O objectivo das decisões estratégicas resume-se ao cumprimento dos principais objectivos e metas traçadas pela organização (e.g. a criação de vantagens competitivas face às suas congéneres). Logo, a qualidade destas decisões pode produzir um impacto global sobre toda a organização como seja pelo incremento da quota de mercado, pela melhoria da satisfação dos clientes ou pela diminuição dos custos. Ainda, na tomada de decisão podem ser percepcionadas três variáveis moderadoras capazes de influenciar a forma como os decisores usam a informação: o excesso de informação, o nível de experiência dos decisores e as restrições temporais [Fischer & Kingma, 2001]. Em suma, as contingências quotidianas vividas pelas organizações conduzem a uma gestão eficaz e eficiente dos dados possuídos e ao tratamento e disponibilização dos dados de modo adequado aos anseios dos agentes de decisão.

Para que a tomada de decisão seja devidamente executada deve-se possuir o conhecimento sobre quais os dados de suporte necessários [Rascão, 2000]. Por sua vez, os mesmos dados podem servir de suporte a diferentes decisões, o que exige uma correcta e adequada gestão dos dados disponíveis, como seja pela agregação ou detalhe dos dados, os níveis de acesso, a prontidão, entre outros [Ballou & Tayi, 1999]. Assim, a gestão dos dados deverá incluir igualmente a divulgação aos decisores dos níveis de qualidade dos dados fornecidos [Shankaranarayan et al., 2003]. No cumprimento das suas tarefas, os decisores necessitam de um quadro de indicadores informativos que permita compreender a organização e o dinamismo do mercado; a possibilidade de analisar tendências e padrões, a avaliação de alternativas e a utilização dos resultados no aproveitamento de oportunidades. Por isso, é extremamente importante possuir um sistema de suporte à decisão que permita à organização permanecer competitiva [Brackett, 1996]. *Sun Tzu*, pai da estratégia, sobre a realidade militar de à 2000 anos, perspectivava:

*"… eram atributos indispensáveis para a «vitória» a capacidade de previsão, de iniciativa, de manobra e de adaptação a novas circunstâncias (…) a necessidade de fazer a escolha certa no momento certo."* [Tzu, 1994].

A consciência desta realidade levou à emergência de iniciativas no campo das tecnologias da informação que sirvam de suporte à tomada de decisão. É o caso dos SDWs como meios tecnológicos capazes de fornecerem as informações exactas e relevantes, no momento oportuno, com o nível de detalhe e em formato adequado para a tomada de decisão. Portanto, são ferramentas que servem de suporte para alcançar os objectivos estratégicos das organizações [Rascão, 2000]. Os

SDWs vieram assumir um papel de relevo no domínio dos processos de tomada de decisão, assumindo claramente a dianteira, relativamente aos sistemas de suporte à decisão anteriormente disponíveis, como uma plataforma tecnologicamente capaz de disponibilizar um conjunto de meios potenciadores de ir ao encontro das preocupações, necessidades e ambições mais essenciais reveladas pelas organizações, em particular pelos seus agentes de tomada de decisão. Os SDWs são uma ferramenta muito útil para suporte às actividades quotidianas dos agentes de decisão, essencialmente na condução táctica das organizações. Este auxílio revela-se importante, na medida em que torna o processo de tomada decisão mais rápido e efectivo, flexibilizando o acesso a mais dados, melhores e mais bem organizados, e garantindo uma maior confiança aos agentes sobre a credibilidade da informação que disponibiliza. Estes são alguns dos vectores basilares que justificam a integração e a exploração de SDWs no seio das organizações. Neste sentido, um SDW apresenta-se como um conjunto de componentes interrelacionados em vista a disponibilização de informações relevantes e indicadores de desempenho aos consumidores finais. Este conjunto de componentes é composto por recursos humanos e tecnológicos, apoiados em técnicas e meios, que actuam sobre as diferentes fases do DW e são capazes de atingir os objectivos finais a que o DW se propõe [Vassiliadis et al., 1999].

O interesse em implementar SDWs pelas organizações pode ser constatado pelo crescimento previsto para o período entre os anos 1996 e 2002. O grupo de gestão de *Palo Alto* perspectivou o crescimento do mercado de DWs de 10 biliões euros, em 1996, para 120 biliões euros, em 2002. Esta previsão configura uma taxa de crescimento na ordem dos 51% anuais durante os anos referidos [Watson et al., 2002]. Os SDWs, pelas suas especificidades, funcionam como sistemas de sentido único, ou seja, a interacção deste tipo de sistemas com os seus utilizadores desenvolve-se apenas no fornecimento de informações vitais com os meios humanos norteadores da condução do negócio. Os utilizadores apresentam-se, assim, como meros consumidores de informação e conhecimento. Esta característica exige que os dados armazenados, nas suas mais variadas dimensões, apresentem elevado grau de qualidade, pois de outro modo implicará a irrelevância do seu interesse e consequentemente resultará na falta de credibilidade do próprio sistema.

## 3.2 Arquitectura dum SDW

Os dados percorrem um longo percurso desde o instante da criação até ao momento da disponibilização de tendências e de indicadores auxiliares aos gestores de topo para condução das organizações. A aferição sobre o nível de qualidade dos dados nos SDWs mostra-se dificilmente atendível se percepcionarmos o sistema no seu todo. Assim, o entendimento cabal sobre o grau da qualidade dos dados existente no sistema pressupõe a necessária desmontagem deste em compo-

nentes elementares perfeitamente balizados nas suas acções, nas tarefas a realizar e nos repositórios de dados que lhes servem de suporte. Cada objecto constituinte do SDW exige um foco de especial atenção, nomeadamente, entre o cruzamento dos dados existentes nos diversos repositórios dos componentes dos SDWs e a qualidade apresentada desses mesmos dados.

O principal papel de um sistema de processamento analítico dos dados consiste em permitir aceder a visões sobre o mundo real de modo a que as pessoas possam tomar decisões [Orr, 1998]. Assim, na óptica dos dados, os SDWs podem ser descritos como um conjunto de vistas materializadas e encadeadas, que actuam em camadas a partir das fontes de dados até à disponibilização dos dados aos utilizadores finais [Theodoratos & Bouzeghoub, 1999]. A justificação pela materialização de vistas sobre os dados surge da possibilidade em melhorar o desempenho das interrogações aos dados e da redução de carregamentos excessivos destes [Bouzeghoub & Peralta, 2004]. A materialização destas vistas inicia-se no nível primitivo dos dados, isto é, nas fontes de dados, vulgarmente, de natureza heterogénea e geograficamente dispersa. Assim, as fontes de dados estabelecem-se como a camada mais baixa deste sistema.

A camada intermédia do sistema designa-se por ARD e consiste num repositório que abarca os dados em bruto e extraídos das fontes. Os dados provenientes do SO podem ser agrupados em três categorias: os dados históricos do SO; os dados residentes no SO e os dados produzidos diariamente no SO como resultado do processamento de transacções [18]. Todos estes dados irão passar por um conjunto de processos nesta área: operações de limpeza, transformação, preparação, agregação, exclusão, integração e carregamento no DW [Chaudhuri & Dayal, 1997]. É também importante salientar que esta camada representa cerca dos 2/3 dos recursos totais consumidos. Opcionalmente, associada a esta camada, é possível observar um outro local, designado por *Operational Data Store* (ODS), capaz de permitir o acesso aos dados em tratamento por parte dos utilizadores. Este local torna-se particularmente interessante na medida em que disponibiliza dados que ainda não se encontram disponíveis no repositório do DW e também outros dados com maior nível de detalhe ao verificado no repositório do DW. Esta situação revela-se particularmente útil na produção de determinadas análises sobre os dados. Uma vez executadas as operações próprias na ARD, os dados são enviados para a camada seguinte, designada por repositório de dados do DW. Neste local concentram-se os dados capazes enriquecerem os conhecimentos dos consumidores de informação e de os auxiliarem no exercício das suas actividades, em particular aquelas relativas à tomada de decisão. Todavia, os consumidores finais necessitam, normalmente, apenas de dados orientados para o desempenho das suas actividades, originando deste modo, acessos a pequenas parcelas do DW total, que dão origem à última camada desta arquitectura e que congrega os DMs [Fabret et al., 1997]. Por fim, deve existir uma camada de suporte, respei-

tante aos metadados, que armazena as informações sobre as características dos dados, dos processos e das diversas camadas que compõem o SDW [Inmon, 2006a] (figura 3-1).

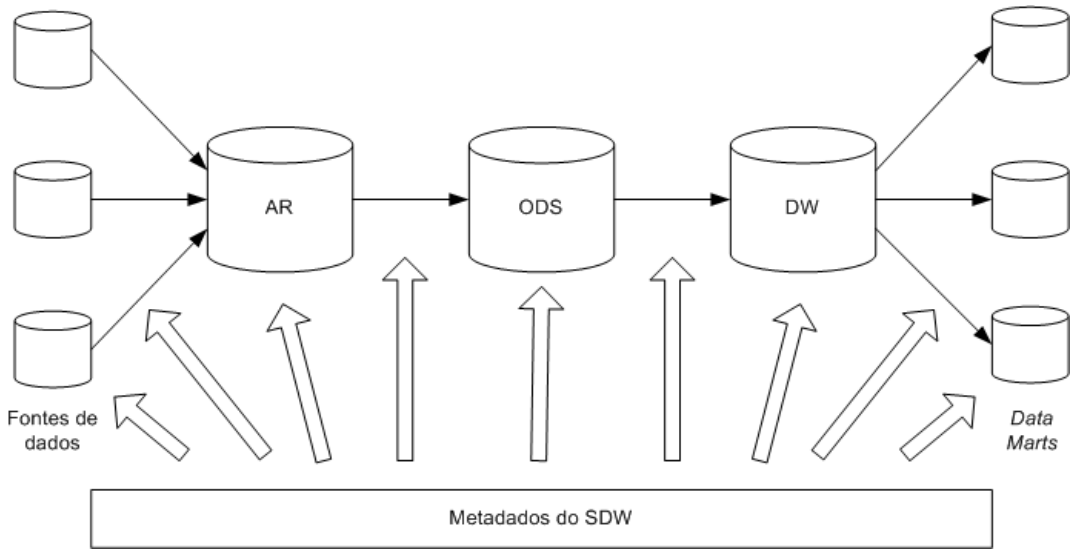



**Figura 3-1 – Arquitectura básica de um SDW.**

Uma arquitectura mais completa pode ser observada em [Bouzeghoub et al., 1999] [Fabret et al., 1997]. Cada bloco da arquitectura tem as suas próprias especificidades, quer em termos de constituintes técnicos, materiais e humanos, quer em termos das incertezas verificadas nos dados armazenados. Esta última questão é o assunto central deste trabalho. Neste pressuposto, é conveniente o esclarecimento num primeiro momento, sobre o impacto e influência da qualidade dos dados que circulam neste tipo de sistemas e num momento posterior, o atendimento sobre o modo como os dados se deslocam durante a sua estada no SDW: desde a entrada dos dados no SO até à disponibilização destes aos consumidores finais.

## 3.3 O impacto da qualidade dos dados num SDW

A implementação de um SDW gera enormes consumos de recursos, desde a recolha dos dados até à disponibilização dos dados aos consumidores finais. Um conjunto de diversas ferramentas informáticas encontra-se presente nas diferentes camadas da arquitectura de um SDW. As organizações começam a colocar na lista de prioridades a aquisição de aplicações que permitam obter dados com elevado grau de pureza, ou no mínimo com o grau de pureza adequado ao uso [Paulson, 2000]. Todavia, estas ferramentas são apenas uma parte do “puzzle” que garante a boa qualidade dos dados. Outras iniciativas de âmbito complementar, têm paralelamente de suportar o processo de melhoria dos dados do DW. Recentemente, a qualidade dos dados tem tomado maior

visibilidade devido a experiências terrivelmente custosas sentidas pelas organizações e por isso, é hoje em dia aceite que a implementação com sucesso de um SDW só possa ser conseguida pela garantia da boa qualidade dos dados [Wang et al., 2003]. Certamente, não será alheia a necessidade de exploração dos dados por ferramentas OLAP e aplicações de mineração. A criação de conhecimento e sabedoria que possibilite o avanço das organizações em teatros de operações fortemente concorrenciais, aproveitando oportunidades e corrigindo fraquezas, determina a existência de um suporte que forneça dados de boa qualidade. Em suma, os dados apresentam-se como um factor crítico de sucesso na construção de um SDW.

### 3.3.1 Custos da fraca qualidade dos dados

O estudo [Watson et al., 2001] realizado sobre um conjunto de organizações norte-americanas é revelador de algumas realidades relativas a estes sistemas. Em primeiro lugar, mostra-nos que os custos intrínsecos à construção dos SDWs rondam em média 1,25 milhões de euros, podendo em certos casos lograr os 47,5 milhões de euros. Os custos operacionais de manutenção anuais atingem em média valores na ordem dos 417 mil euros, podendo alcançar em alguns casos os 7 milhões de euros. Quanto à comparação entre os benefícios esperados e os realmente atingidos, constata-se que os maiores desvios são os relativos à possibilidade dos utilizadores realizarem mais e melhores consultas, à maior rapidez da tomada de decisão, à melhor qualidade de informação e à maior facilidade na tomada duma decisão alternativa. Esta situação é efectivamente constrangedora, uma vez que contraria os propósitos basilares dos SDWs. Por último, este estudo mostra que menos de um sexto das organizações consideram como pleno de sucesso o seu SDW. Idêntica realidade é observada em [English, 2002a], referindo que apenas 8% dos SDWs implementados apresentam sucesso ao fim de 3 anos de operacionalidade, enquanto outro estudo [Watson et al., 2002] aponta para uma taxa de insucesso dos SDWs em torno dos 50%. Este cenário é obviamente preocupante na medida em que possibilita aquilatar sobre a extrema dificuldade em construir e manter um sistema deste porte.

Outros estudos são reveladores do facto da grande maioria das organizações considerarem a qualidade dos dados como o principal factor de sucesso dos SDWs. É o caso do estudo promovido pelo *Metagroup* [Dataflux, 1999], que inquiriu cerca de 3 mil utilizadores de SDWs, os quais consideram a qualidade dos dados como o principal desafio a enfrentar em ambientes de suporte à decisão. Este estudo refere ainda que, segundo dados obtidos pelo *Data Warehousing Institute*, um quarto das empresas norte-americanas promovem iniciativas de melhoramento dos seus dados e que cerca de 15% dos dados relativos aos seus clientes estão incorrectos. O mesmo instituto estima que os custos associados à fraca qualidade dos dados se situam em torno dos 500 bi-

liões de euros anuais e que o problema principal prende-se pelo facto dos gestores de negócio não possuírem a perfeita consciência desta situação. Este estudo vai ainda mais longe ao afirmar que a degeneração da qualidade dos dados está lentamente sangrando as empresas até à morte [Eckerson, 2002]. Os relatos de casos de insucesso no domínio dos SDWs sucedem-se e explicam-se, em última instância, por razões da qualidade dos dados nos sistemas. O fracasso na implementação do SDW dum grande banco, que custou cerca de 30 milhões de euros ou o fiasco dum SDW governamental, avaliado em 21 milhões de euros são entendidos como originados pela menor atenção prestada à qualidade dos dados existentes [English, 2002a].

As dificuldades sentidas ao nível da captação de dados apropriados para constarem num DW são uma realidade e um obstáculo por vezes difícil de vencer. Alguns estudos referem que os projectos de migração dos dados ultrapassam rapidamente o orçamento inicial previsto como resultado de maus entendimentos sobre as fontes e das definições dos dados, sendo apontadas as deficiências nos dados como a principal razão para furar os orçamentos e falhanços dos projectos [Igor & Mahnic, 2000]. Em [Orr, 1998] faz-se referência a um gestor que aponta para o facto de cerca de 60% dos dados que circulam do SO para o DW falharem quando são impostas as regras do negócio. Um estudo [Neely, 1998] refere um artigo publicado no *Wall Street Journal*, de 1998, que relata o efeito dominó consequente quando dados erróneos tipificam a base de dados duma organização. Um outro estudo promovido pelo *Metagroup* revela que 41% dos projectos de DW falham redondamente e a principal causa identificada é a fraca qualidade dos dados constantes no DW [Jarke et al., 2003].

A pouca qualidade dos dados significa que as informações apresentadas apresentam deficiências numa ou mais vertentes ou dimensões dos dados. Esta situação estabelece o princípio *garbage in, garbage out*, em que os maus dados conduzem a más decisões, que se traduzem em insatisfações dos clientes, em perdas de oportunidades de mercado, em dificuldades na escolha sobre a melhor decisão entre várias alternativas e em tomada de decisões estratégicas erróneas ou equivocadas. Assim, a qualidade dos dados assume-se como um factor chave para o sucesso dum SDW [Jarke et al., 2003]. As organizações vivem e morrem em função da inteligência que conseguem imprimir no desenho da qualidade dos seus dados. A inteligência provem da combinação concertada de meios técnicos e recursos tecnológicos, como é o caso dos SDWs. Porém, a inteligência só é boa se a qualidade dos dados for óptima [Nguyen & Fisher, 2000]. Em suma, perante um panorama caracterizado por elevados índices de insucesso numa parte significativa dos SDWs organizacionais, pode-se assumir que a implementação com sucesso dos SDWs se encontra associada a investimentos de alto risco, exigindo o envolvimento da organização, tendo em vista encontrar soluções que minimizem os riscos inerentes a estes sistemas.

### 3.3.2 Benefícios da qualidade dos dados

A determinação dos benefícios obtidos pela implementação de um SDW mostra-se, normalmente, uma tarefa de difícil execução, uma vez que esses benefícios, especialmente os mais valorizados, apresentam impactos intangíveis. Além disso, o tempo necessário para a execução desta tarefa é geralmente longo, rondando em média os 3 meses de duração [Watson et al., 2002]. Apesar destas dificuldades, algumas técnicas e metodologias têm surgido ao longo dos anos no sentido da determinação dos benefícios. Os benefícios facultados pelo desenvolvimento de DWs podem classificar-se em tangíveis e intangíveis. Algumas investigações identificam cerca de cem benefícios possíveis de alcançar, como sejam: a precisão dos dados, a facilidade de utilização, o tempo de resposta, a informação útil, entre muitos outros. A natureza dos principais benefícios verificados apresenta-se como um obstáculo na sua enunciação e posterior quantificação. Dada a vasta amplitude de benefícios, o agrupamento destes por níveis permite um tratamento mais sintético e consequentemente uma gestão mais manuseável e viável.

No estudo [Watson et al., 2002] é desmistificada a tangibilidade dos benefícios segundo dois eixos: o impacto verificado e a facilidade de medição. O primeiro eixo proposto pretende reconhecer a abrangência provocada pelo impacto do benefício na organização, isto é, se gera um efeito localizado ou se pelo contrário produz consequências na globalidade da organização. Quanto ao segundo eixo, o objectivo passa por averiguar a maior ou menor facilidade ou dificuldade na medição dos benefícios. A visualização dos benefícios segundo a modalidade proposta permite a identificação dos benefícios tangíveis, como os mais fáceis de medir por terem um impacto localizado no seio da organização e dos benefícios intangíveis, que traduzem maiores dificuldades na sua medição porque o âmbito do impacto é reflectido em toda a extensão da organização (figura 3-2). Estes últimos, representam os melhores benefícios que um DW permite auferir, porque surgem quando o DW é utilizado no redesenho dos processos do negócio e no suporte à concretização dos objectivos estratégicos do negócio.

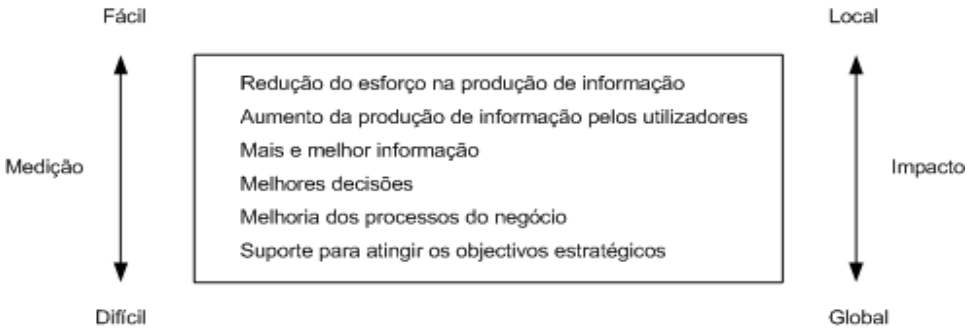



**Figura 3-2 – Benefícios do DW [Watson et al., 2002].**

A percepção sobre os benefícios a atingir, derivados da utilização dum DW, pressupõe que os valores dos dados constantes no sistema se apresentem num estado de qualidade compatível com os anseios da organização. Naturalmente, a gestão criteriosa dos recursos, certamente limitados, das organizações pode implicar que a obtenção da qualidade dos dados mais desejada não seja satisfeita porque o esforço financeiro necessário não compensa os resultados obtidos. Logo, os investimentos realizados na tentativa de atribuir maior qualidade aos dados devem ser perspectivados sobre o ROI a obter. A realização de um estudo comparativo das diferentes possibilidades de investimento a realizar e consequente, retorno a obter, permitirá optar pela solução mais conveniente para a organização [McKnight, 2003]. A tabela 3-1 ilustra algumas possibilidades de escolha dum nível de qualidade dos dados apropriado.

| Nível de qualidade dos dados | Investimento | ROI |
|---|---|---|
| 90 | € 400 000 | 175% |
| 85 | € 390 000 | 101% |
| 80 | € 380 000 | 65% |

**Tabela 3-1 – Análise custo e benefício [McKnight, 2003].**

A conversão dos benefícios em termos de ROI realizado em SDWs é geralmente de difícil concretização devido, conforme descrito anteriormente, à morosidade de cálculo dos benefícios e às dificuldades verificadas na enunciação e quantificação dos mesmos. Esta realidade é observável no estudo [Watson et al., 2001], desenvolvido sobre SDWs, em que apenas 1/4 dos inquiridos responderam sobre a questão relativa ao ROI alcançado. Porém, as organizações que responderam sobre esta questão referiram, em média, valores em torno dos 300%. Um outro estudo promovido por [Watson et al., 2002] aponta para a obtenção de valores de ROI em torno dos 400% a 3 anos. Em suma, desde que bem implementados estes sistemas produzem efeitos claramente positivos [Kenyon et al., 2004].

## 3.4 As razões da fraca qualidade dos dados em SDWs

Actualmente, assiste-se a um contínuo crescimento do volume e da complexidade dos dados existentes nas organizações. Esta situação conduz ao crescimento da importância da qualidade dos dados presentes nos sistemas organizacionais porque os mesmos dados são a base para informações distintas, a sua distribuição por diferentes locais revela-se uma urgência e pelo facto de denotarem um carácter de transversalidade em todas as actividades organizacionais [Brackett, 1996]. Os dados assumem-se como a matéria-prima na era da informação [Ballou & Tayi, 1998] – tal qual os ovos numa omeleta. Contextualizando ao domínio dos SDWs, podemos encarar as

razões da fraca qualidade dos dados como desafios a enfrentar e que urge resolver no sentido de alcançar os objectivos primeiros deste tipo de sistemas [Strong et al., 1997].

### 3.4.1 Natureza estratégica

Posicionando-nos no topo da hierarquia organizativa, verifica-se a absoluta ou parcial inconsciência ou displicência sobre os dados que povoam as organizações. Em [Redman, 1998] alerta-se para o facto de tipicamente os gestores estarem já ocupados com demasiados problemas a nível organizacional, como seja a insatisfação dos clientes, os custos elevados na implementação das infra-estruturas tecnológicas, os atrasos temporais no projecto de DW, entre outros e por isso, a pouca atenção prestada à qualidade dos dados é uma consequência lógica. Todavia, apesar de inicialmente negligenciada, este assunto tem demonstrado ser um tema central na vida das organizações. Actualmente, é um dos aspectos mais importantes que os gestores enfrentam porque se trata da questão recalcada da origem de muitos problemas ao nível organizacional, em termos gerais e na implementação e manutenção dos SDWs, em termos específicos. Em [Ballou & Tayi, 1998] é corroborada a opinião de *Redman* sobre o alheamento das organizações face aos dados e à qualidade que estes apresentam. Esta situação traduz-se em dificuldades no plano operacional da implementação dos SDWs, como sejam, basicamente, a combinação de dados que não foram projectados para serem integrados, o armazenamento de dados impróprios e a baixa prioridade em assegurar a qualidade dos dados. Estas dificuldades ofuscam e muitas vezes inviabilizam a compreensão sobre a natureza das deficiências dos dados. Citando [Ballou & Tayi, 1999]: *"Ignorar ou relativizar os problemas com os dados existentes, no início do projecto, é deixar que os erros venham a revelar-se brutalmente a si próprios..."*.

Um estudo da consultora *PriceWaterhouseCoopers* [Kenyon et al., 2000], confirma algumas das razões da fraca qualidade dos dados anteriormente invocadas e motivadas pela gestão de topo, como sejam: o pouco conhecimento dos problemas e dos custos da fraca qualidade dos dados; a complacência dos administradores das organizações quanto às questões relativas aos dados; a delegação de responsabilidades no lugar errado; as reacções aos poucos, sem medidas estratégicas e a desconfiança interna e externa na qualidade dos dados da organização. O estudo situa estas questões ao nível das acções estratégicas das organizações e sintetiza que a gestão dos dados está erradamente endereçada ao nível errado, no lugar errado e na forma errada. O mais recente estudo da mesma consultora, reforça a necessidade do envolvimento de toda a organização e uma liderança forte na acção, tratando a gestão dos dados como um factor crítico de sucesso dos SDWs e consequentemente no futuro das organizações, qualquer que seja a sua natureza [Kenyon et al., 2004].

### 3.4.2 Natureza operacional

As razões da fraca qualidade dos dados podem ser observadas no âmbito dos motivos operacionais responsáveis pela fraca qualidade dos dados num SDW. Assim, para um melhor entendimento das causas dos problemas, é possível cruzar essas causas com os diferentes componentes constituintes da arquitectura dum SDW: o SO, a ARD, o repositório do DW e os DMs.

#### Sistema operacional

O SO assume importância vital nos SDWs porque são a nascente do fluxo circulatório dos dados que percorrem todo o sistema. Como tal, um factor que pode implicar a falta de qualidade dos dados prende-se com os dados existentes ou não nas fontes de informação, isto é, não podemos exigir ao sistema respostas sobre dados que não possui, ou que é impossível obter. Ou de modo mais ténue, sobre dados vincados por falhas de qualidade irrecuperáveis nas diversas propriedades que os caracterizam: consistência, acessibilidade, validade, exactidão e relevância.

A recolha e o armazenamento de dados sujos no SO ocorre devido a um leque alargado de questões, como sejam, segundo [Inmon et al., 1998]: a elementos opcionais no *software* de aquisição de dados, ao facto dos elementos de *software* serem definidos duma forma e o programador decidir doutro modo e a especificações incorrectas desde o princípio. Ainda neste âmbito, em [English, 1999] é realçada a pouca integração dos dados dispersos e redundantes, a aceitação de valores falsos, geralmente, dificilmente reparáveis, a inexistência de uma análise de requisitos que satisfaça os anseios dos utilizadores e a proliferação e variedade de erros nos dados no SO das organizações. Este é um tema particularmente caro para *Olson*, pois considera que os processos de tratamento dos dados no SO, em especial, aqueles relacionados com a aquisição dos dados condicionam as futuras utilizações dos mesmos [Olson, 2003].

#### Área de retenção dos dados

Normalmente, a preocupação com a qualidade dos dados existentes no SO ocorre no momento destes serem usados no DW. Neste instante, várias opções podem surgir. A primeira, segundo *English*, constitui um dos erros típicos na implementação de SDWs e passa pela assumpção que as fontes de dados são boas porque o SO funciona bem [English, 2002a]. Uma segunda opção, mais consciente, poderá passar pelo abandono do projecto devido à proliferação de dados irregulares no SO que inviabiliza a construção de um SDW fiável e possuidor das características adequadas ao uso [Kimball & Caserta, 2004]. Por último, na opção mais comum, os problemas de qualidade dos dados tendem a ser melhor ou pior resolvidos na ARD afim de obter um repositório de dados de melhor qualidade. Este local deverá ser capacitado dos processos necessários para

resolver os problemas relativos aos dados provenientes da diversidade de fontes dispersas e de natureza heterogénea, as incompatibilidades ao nível da estrutura de chaves, da estrutura dos dados, da codificação dos dados, da definição dos dados, da detecção de valores duplicados, das características físicas dos dados, entre outras [Inmon et al., 1998]. Assim, os problemas com sujidade dos dados são tratados e estes uma vez rectificados são carregados durante o primeiro carregamento dos dados no DW.

Nos carregamentos subsequentes dos dados, os cuidados e rectificações tendem a ser análogos aos ocorridos no primeiro carregamento. Usualmente, os carregamentos posteriores não resultam de reposições integrais da totalidade dos dados, mas antes de incrementos aos dados já existentes e que se designam por carregamentos incrementais. Esta opção é justificada pela redução do volume de dados a incorporar num DW e pelo tempo disponível concedido durante a janela de oportunidade [Chaudhuri & Dayal, 1997]. Assim, a exigência de uma maior prontidão das informações, pelos consumidores finais do sistema, poderá implicar uma maior rapidez deste processo, que ocorre durante uma janela de oportunidade temporal. Porém, a janela de oportunidade poderá não ser suficientemente ampla para albergar integralmente todo o processo de ETL dos dados e em consequência originar a ocorrência de dados imperfeitos ou incompletos no DW. Ora, os consumidores finais podem optar pela rapidez ou presença de dados frescos em detrimento da perfeição dos dados, pois caso contrário alguns dados poderiam nunca se encontrar disponíveis no momento da tomada das decisões. Na verdade, muitos utilizadores do DW necessitam menos do que dados perfeitos para efectuar as análises, estatísticas e agregações sobre os dados [Cappiello et al., 2004]. Na óptica da tomada de decisão o uso da melhor informação poderá não ser a mais completa o que pode viabilizar a existência de dados não perfeitos no sistema [PMBok, 2000]. Assim, a premência na obtenção de informações apresenta-se como um outro factor para a existência de dados imperfeitos [18]. Assim, a focalização dos recursos em vista a melhoria do desempenho em vez da garantia de uma melhor qualidade dos dados revela-se como mais um erro fulcral na implementação de SDWs [English, 2002a]. Na prática, o desequilíbrio entre as dimensões dos dados pode originar falhas na qualidade destes.

Neste contexto, verificamos que a ARD se mostra o local, por excelência, guardião da qualidade dos dados, mas igualmente, uma zona susceptível de gerar anomalias nestes porque existem muitos detalhes a merecerem a necessária ponderação de forma a afiançar o cumprimento eficaz e eficiente das tarefas, dos constrangimentos e dos critérios a respeitar. A objectividade e clareza das actividades que envolvem o processo de ETL são um factor condicionante para a obtenção dos dados de acordo com os requisitos definidos [Kimball & Caserta, 2004]. A falha ou debilidade num processo de tratamento dos dados pode ser a causa da existência de problemas a nível da

qualidade dos dados e consequentemente, provocar um impacto negativo na organização. Mesmo considerando que estas questões se encontram em vias de resolução, outros contratempos podem surgir que abanam a estrutura de garantia da qualidade dos dados existente, como seja a adição de novas fontes de dados ao SO, a substituição de algumas fontes de dados existentes ou a introdução de dados não estruturados (e.g. mensagens de *email*, imagens, etc.) [Inmon, 2006b]. A ocorrência destas contingências pode comprometer que alguns dos processos deste local se tornem incompletos, ultrapassados ou incorrectos (e.g. regras de transformação obsoletas ou a alteração das regras do negócio).

## Repositório do DW

A imperfeição dos dados contidos num DW deve-se, essencialmente, à circunstância dos dados existentes no repositório assentarem em dados históricos, isto é, os dados existentes no sistema reportam-se a períodos de tempo alargados. Esta situação pode provocar a inoperacionalidade do valor dos próprios dados, em virtude do enquadramento temporal ser outro e das regras ou requisitos de negócio terem-se modificado [18]. Deste modo, apesar das condições do negócio serem outras, os dados introduzidos anteriormente mantêm-se inalterados. Mesmo considerando que o SO se apresenta completamente limpo e que os processos de integração e transformação sejam considerados perfeitos (o que não é verdade), continuarão a existir dados sujos no DW, devido à idade e consequente desactualização destes no próprio sistema [Olson, 2003]. Portanto, os dados deterioram-se com o tempo (e.g. fusões ou separações de organizações, a mudança de sistema informático, a mudança do próprio negócio ou os dados pessoais e de gestão dos clientes).

## Data marts

Os DMs são a camada directamente em contacto com os consumidores e por isso podem revelar questões particularmente críticas em relação aos dados. Estas questões podem-se perspectivar como anomalias de interpretação, credibilidade e utilidade dos dados apresentados e que decorrem da origem semântica dos dados e da multiplicidade de consumidores existentes. Mesmo considerando-se, por hipótese, a correcção dos dados, a variedade de consumidores, certamente, origina diferentes interpretações sobre estes. Igualmente problemática revela-se o desprovimento da capacidade de julgamento sobre razoabilidade dos dados por parte dos consumidores, uma vez que estes não possuem qualquer responsabilidade pela integridade dos dados. Daí a importância da credibilidade das fontes de dados presentes no sistema [Ballou & Tayi, 1998].

Ainda neste local é necessário salientar o dinamismo associado ao negócio das organizações e consequentemente às plataformas tecnológicas que lhe servem de suporte porque podem exigir a mudança dos requisitos dos consumidores finais ou a entrada de novos consumidores de informa-

ção no sistema. Pressupondo que um SDW não possui as capacidades para estas novas exigências da organização a reclamação de uma manutenção adequada e precisa que vise a introdução de novas perspectivas sobre os dados e a qualidade dos mesmos torna-se fundamental.

### 3.4.3 Análise de consequências

As razões anteriormente invocadas permitem retirar algumas ilações em vista garantir a efectiva qualidade dos dados. A primeira refere-se aos custos inerentes à presença de dados fracos no sistema. Estes custos devem ser ponderados com os custos associados à não disponibilização dos dados aos utilizadores em tempo considerado útil ou à divulgação de dados deficientes. Ainda, relativamente aos custos importa salientar que estes não se reportam somente aos custos declarados, mas compreendem também os custos ocultos. Deste modo, o custo real, resultante da adição dos custos declarados e dos custos ocultos, apesar da impossibilidade prática na sua determinação, deve ser uma componente a considerar em relação ao nível de qualidade dos dados exigida.

Uma outra ilação reporta-se à incapacidade em possuir apenas dados perfeitos nos SDWs, o que antecipa e prevê a possibilidade de serem estabelecidos níveis de imperfeição sobre os dados. A imperfeição dos dados não significa a inutilidade desses dados, ou seja, alguns dados tornam-se mortos quando incorrectos, enquanto que outros apenas podem ser considerados imperfeitos, quando incorrectos. Recorrendo a um exemplo apresentado por *Inmon*, em que é colocada a questão em saber se uma tabela com um milhão de linhas que contem uma linha incorrecta e uma outra que contem dez linhas, em que nove são inválidas, podem ou não representar fraca qualidade dos dados. Perante o dilema em definir o nível de qualidade dos dados, é defendido que a resposta depende da natureza de cada tabela considerada [18]. Assim, o administrador do DW deve definir a importância sobre cada coluna de dados, porque é possível que uma coluna necessite de se apresentar completamente limpa e integrada, enquanto uma outra coluna pode não respeitar os padrões máximos de perfeição.

Decorre, destas duas ilações, uma outra que visa a ponderação sobre a necessidade em defender uma manutenção da qualidade dos dados que tenha em linha de conta a participação das organizações num mercado cada vez mais voraz e eliminatório e que exerce uma pressão contundente no sentido destas repostarem mais rapidamente às solicitações internas e externas. Perante tamanha exigência as organizações sentem a necessidade de possuírem SDWs mais próximos das exigências quotidianas, ou seja, tendentes ao *real-time*. Esta conjuntura gera a necessidade em capacitar a organização para a promoção da mudança organizacional, que envolva todos os inter-

venientes na organização e a redefinição de conceitos, estruturas funcionais, estruturas tecnológicas, planos de formação, entre muitas outras [Watson et al., 2002].

## 3.5 Os problemas da qualidade dos dados num SDW

A arquitectura de um SDW constitui-se em quatro camadas essenciais: o SO, a ARD, o repositório de dados do DW e os DMs que interagem com os utilizadores. À luz do defendido por *Inmon* [18], cada uma destas camadas é fonte de um conjunto de problemas relativos à qualidade dos dados e deve merecer especial atenção em vista o seu melhoramento. Esta visão pode ser directamente vertida nos quatro motivos principais para a existência de dados sujos nos sistemas informáticos defendidos em [Olson, 2003]: a entrada incorrecta dos dados, o transporte e reestruturação dos dados, a decadência dos dados e a utilização dos dados. Assim, interessa perceber quais os tipos de problemas associados a cada camada em vista a sua identificação, entendimento, classificação e possível método de rectificação [Helfert, 2003].

### 3.5.1 Qualidade dos dados no SO

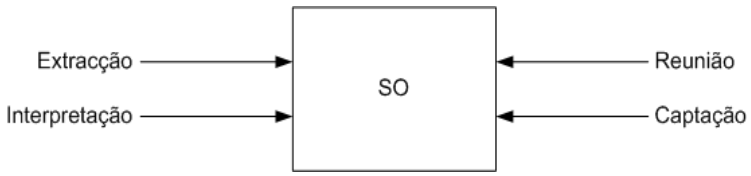



**Figura 3-3 – Problemas nos dados no SO.**

O problema central a nível do SO consiste, regra geral, no desconhecimento da veracidade dos dados existentes e dos níveis de qualidade que estes apresentam. As exigências ou requisitos dos consumidores podem não ser correspondidas adequadamente porque os dados disponíveis nas fontes são escassos ou apresentam problemas dificilmente reparáveis. A ponderação sobre a efectiva implementação dum SDW deve sempre levar em linha de conta se existe matéria-prima de boa qualidade na produção de informação. Neste sentido, o conhecimento sobre todas as características das fontes que compõem o SO deve ser realizada. Esta questão faz ressaltar o aparecimento de novas ferramentas, como sejam as técnicas de análise dos dados (e.g. *data profiling*), tendo em vista a compreensão e apresentação de modo mais fiável das características das fontes de dados [White, 2000]. Em [Kimball & Caserta, 2004] é exultado o interesse na utilização deste tipo de técnicas e é explicitado abertamente a opção pela desistência de concretização do SDW, caso os dados do SO não apresente os níveis exigíveis para tal realização. Portanto, a garantia da boa qualidade dos dados no SO promove as bases para uma boa qualidade dos dados

nas camadas do SDW subsequentes. Por outras palavras, caso os dados constantes na camada inicial não sejam fiáveis, então a circulação dos dados no SDW dificilmente será de confiança. Ora, a existência de SOs perfeitamente limpos é praticamente impraticável e certamente dispensável de garantir. A natureza das fontes de informação é, geralmente, heterogénea e dispersa, isto é, os sistemas alvos da extracção dos dados apresentam-se estruturalmente muito distintos entre si. Acresce a possibilidade de integração de dados no DW provenientes de fontes internas e externas à organização, com fusos horários e localizações geográficas profundamente dispersas, denotando níveis de consistência, completude e exactidão duvidosas, provenientes de diferentes épocas temporais e apresentando os mais variados formatos (folhas de cálculo, sítios da *web*, imagens, colunas sobrecarregadas de valores, ficheiros de texto, base de dados hierárquicas, base de dados relacionais e até mesmo em suporte papel) (figura 3-3).

No domínio do SO, mesmo considerando as aplicações de captura de dados mais recentes, a introdução de dados continua a ser o principal foco de origem das debilidades da qualidade dos dados devido a diversos factores, como sejam [18] [Dataflux, 1999]:

- Aplicações de recolha de dados complexas e que induzem a inserção de erros (e.g. um campo que não especifica se deve ser inserido local de nascimento ou de residência).

- Dificuldades na inserção completa e correcta dos dados por limitações temporais na recolha dos dados ou pelo facto dos funcionários não estarem devidamente consciencializados e motivados para a importância da inserção de dados correctos e completos.

- Inserções involuntárias de dados válidos mas incorrectos (e.g. o operador digita a idade igual a 43 em vez de 34) ou voluntária de valores errados (e.g. inserção sistemática de valores por defeito ou valores fictícios).

Assim, a existência de erros nas bases de dados é uma consequência factual da vida das organizações e pode ser profundamente perturbadora na implementação e manutenção com sucesso dos SDWs. A inserção incorrecta de valores, de forma deliberada ou não, pode condicionar o sucesso de um SDW, devido à existência da corrupção de valores de dados, possivelmente vitais, na base de dados. Ou pelo menos, na melhor das hipóteses fará subir os custos associados à transformação dos dados em vista a sua limpeza e adequação ao uso [Neely, 1998] [Novabase, 2002]. Este problema é agravado porque muitos dados apenas se encontram disponíveis em instantes pontuais ou de rara oportunidade de captação (e.g. os sinais vitais de um paciente) e a não inserção nessa ocasião inviabiliza a sua captação em momentos posteriores. A descoberta de soluções que implementem processos automáticos de introdução de dados é apontada como uma iniciativa

a optar [18]. Todavia, os SDWs não consistem somente em aplicações recentes possuidoras de características que assegurem a qualidade dos dados nas suas mais variadas vertentes, mas antes fundeiam-se numa diversidade de fontes de dados constantes nos SO, que compreendem algumas aplicações antigas, outras pobres em termos de consistência e integração dos dados, cuja a limpeza integral dos dados se torna irrealizável. Ainda segundo *Inmon*, se esperarmos que o SO apresente somente dados limpos, então o DW nunca será construído [18].

Em [Gonzales, 2004] é advertido para a determinação das melhores fontes de dados porque, por um lado, existem várias fontes que contêm os mesmos elementos e por outro lado, existem fontes que têm necessidade de serem complementadas com outras fontes (e.g. diferentes níveis históricos dos dados). Esta problemática complica-se ainda mais pela dificuldade na compreensão das próprias fontes de dados. Em especial, no que concerne à complexidade associada na movimentação de dados para e de outros sistemas (e.g. reconciliação de bases de dados, processos de fusão das organizações, substituição de sistemas informáticos), a aceitação de valores de qualidade inferior nos sistemas e o entendimento estrutural das fontes (e.g. a existência num mesmo campo de dois tipos de valores, símbolos com significado especial). Assim, aliado aos problemas existentes, a pouca atenção prestada na manutenção de dados correctos e completos nos repositórios dos metadados do SO são problemas verificáveis e geradores de entraves à implementação com sucesso do SDW [Olson, 2003]. Outro tipo de problemas verificável no SO consiste no modo como os dados são reflectidos num DW. A definição da política de refrescamento dos dados adequada às exigências dos consumidores tem necessariamente de ser enquadrada de modo, a que os dados disponibilizados mostrem padrões de actualidade compatíveis à tomada de decisões.

Neste contexto, podemos inferir que o ambiente operacional é por excelência o local privilegiado para introduzir normas e princípios de qualidade sobre os dados. Este assunto tem assumido particular destaque dadas as exigências de SDWs cada vez mais capazes e dotados de melhores características em termos do desempenho das consultas, segurança de acesso e ao nível da qualidade dos dados apresentada. Esta contingência corresponde às preocupações avançadas em [Kimball & Caserta, 2004] ao ser considerado que o tratamento e limpeza dos dados deve ser concretizado, na maioria das suas tarefas, no SO. Porém, a consciência dominante alerta para a pouca margem de manobra do administrador do DW em agir sobre o ambiente operacional, porque este assenta em aplicações e sistemas de dados enraizados e consolidados pela estrutura informativa da organização. Assim, a manutenção duma auditoria preventiva e correctiva relativa à qualidade dos dados nas fontes, poder-se-á apresentar como um modo híbrido de antecipar e tratar alguns problemas existentes para que os esforços na construção dum DW possam ser melhor correspondidos, enquanto a estrutura funcional e organizativa permanecer inflexível.

### 3.5.2 Qualidade dos dados na ARD


**Figura 3-4 – Problemas dos dados na ARD.**

A ARD corresponde ao local que medeia a passagem dos dados entre as fontes e o DW. Este local apresenta-se como um ponto óptimo para impor normas e níveis de qualidade aos dados provenientes das diferentes fontes. Efectivamente, revela-se o local propício para as operações de limpeza, transformação e combinação dos dados, antes destes serem integrados no DW, ou seja, concentra todas as tarefas relativas ao processo de ETL dos dados no DW (figura 3-4). É por isso, o lugar onde as organizações têm a possibilidade de encontrar o maior nível de qualidade sobre os dados. Segundo *Inmon*, a importância da ARD deve-se a dois factores [18]. Em primeiro lugar, devido à inexistência ou escassez de iniciativas de promoção da melhoria da qualidade dos dados no SO. Esta situação, justifica-se por um lado, pela não interferência dos responsáveis do DW nos ambientes operacionais em vista a introdução de mecanismos promotores do aperfeiçoamento dos dados existentes e que servirão de base para o DW. O princípio clássico de implementação de um SDW delimita as fronteiras e a área de actuação entre o SO e o DW. Por outro lado, a assumpção que esta é uma tarefa incumbente da ARD.

Em segundo lugar, a circunstância de existirem múltiplas fontes para os mesmos dados ou que permitem a complementaridade entre os dados, determina que seja este, vulgarmente, o primeiro local de encontro desses dados dispersos. A definição das fontes a serem usadas é uma tarefa que envolve alguma complexidade, porque a existência de várias fontes para os mesmos elementos de dados determina a escolha daquela mais apropriada e que melhor satisfaz os desejos dos consumidores. Também, o facto das fontes manifestarem diferentes níveis históricos dos dados, pode implicar o recurso a outras fontes de modo a complementar os dados em falta [Gonzales, 2004]. Determinadas questões relacionadas com os dados apenas podem ser resolvidas no momento após a sua extracção das fontes e o recolhimento na ARD (e.g. os dados provenientes de sistemas independentes e complementares) [Kimball & Caserta, 2004]. Assim, os dados extraídos do SO e recolhidos na ARD, denotam lacunas no seu conteúdo, formato e integração que necessitam ser colmatadas. É preciso ter presente que os níveis de qualidade inerentes aos dados variam em função dos requisitos dos consumidores e pela limitação orçamental imposta pela própria organização. Sobre esta última questão é plausível que a organização considere injustificado o investimento para elevar ao máximo os padrões de excelência da qualidade dos dados.

Geralmente, os dados provenientes do SO apresentam fraca qualidade e por isso, costumam designar-se por dados sujos. Os dados consideram-se sujos porque compreendem uma variedade de defeitos (ruídos, inconsistências e dados incompletos). Os dados incompletos revelam a ausência de valores, a ausência de atributos de interesse ou respeitam somente a valores agregados. Os dados manifestam ruído quando são introduzidos valores aberrantes, erros aleatórios e erros provocados pelo sistema. A inconsistência dos dados surge da introdução incorrecta de valores, de representações diferentes sobre os mesmos dados ou de erros oriundos da integração de várias bases de dados [Carvalho, 2003]. Em [Raisinghani, 1999] é enunciada uma lista que tipifica a sujidade dos dados em ambientes de DW, como sejam a ausência, a incorrecção, a incompreensibilidade, a inconsistência, os conflitos de esquemas e os conflitos estruturais dos dados. Um estudo realizado em [Kim et al., 2003], aprofunda esta problemática em torno dos dados e concretiza uma taxionomia exaustiva sobre os dados sujos. O estudo aborda este assunto, partindo da premissa que os problemas dos dados se podem apresentar sob três diferentes formas: os ausentes; os não ausentes, mas errados e os não ausentes e correctos, mas inúteis. Com base nesta premissa é possível decompor a taxionomia hierarquicamente, representando cada nível de decomposição ou refinamento uma forma especifica de sujidade nos dados. Ao mesmo tempo estabelece as bases para a definição de métricas que avaliem a qualidade dos dados, bem como as técnicas adequadas de rectificação. A investigação realizada em [Rahm & Do, 2000] descreve uma classificação dos problemas nos dados, consoante se trate de problemas numa fonte de dados isolada ou de problemas que se revelem em múltiplas fontes de dados. Um outro estudo, relacionado com os dois anteriores e assente nos estudos realizados em [Oliveira et al., 2004, 2005b], aborda mais exaustivamente a identificação, classificação e sistematização dos problemas verificáveis na qualidade nos dados. O estudo apresenta uma taxionomia organizada em diferentes níveis de abstracção: em termos absolutos do valor das colunas numa tabela (ao nível da coluna, ao nível da linha e ao nível da tabela) e em termos relativos, resultantes do relacionamento entre tabelas [Oliveira et al., 2005a].

Neste sentido, podemos aferir que o deslocamento dos dados entre o SO e o DW não resulta duma simples extracção dos dados. Mas, de um conjunto de tarefas de cariz complexo e moroso em vista a integração dos dados no DW e que, por estes motivos, são necessariamente responsáveis pela maior fatia no consumo dos recursos financeiros e temporais. Assim, os dados na ARD estão sujeitos à aplicação de diversos processos de limpeza e transformação, com o intuito da remoção das debilidades verificadas nos dados e a melhoria da sua qualidade. A limpeza dos dados, segundo o estudo realizado em [Müller & Freytag, 2002], consiste num processo semi-automático que compreende a execução das seguintes operações sobre os dados: a adaptação dos formatos

das linhas das tabelas e valores, o cumprimento dos constrangimentos de integridade, a derivação de valores em falta a partir dos existentes, a remoção de conflitos nas linhas ou entre linhas, a junção ou eliminação de duplicados e a detecção de desvios, isto é, as linhas que apresentam alta probabilidade de serem inválidas.

Portanto, para que a operacionalidade e utilidade dos dados seja conseguida, por parte dos consumidores finais, importa efectuar o necessário enquadramento das acções de transformação sobre os dados sujos extraídos das fontes. As operações de limpeza e transformação a serem executadas neste local podem ser divididas em duas tarefas fundamentais: o tratamento dos dados e a preparação dos dados para os consumidores finais.

## Tratamento dos dados

No que concerne aos dados recolhidos das fontes e que denotam anomalias nas suas características, podemos considerar cinco categorias de problemas que conduzem a operações de resolução [Kimball et al., 1998] [Raisinghani, 1999]: a decomposição dos valores em elementos atómicos, a estandardização e normalização de valores, a correcção de dados, a integração de dados e a remoção de dados. A decomposição dos valores dos dados em elementos atómicos ocorre:

- Pela existência de colunas sobrecarregadas de valores ou colunas de formato livre.
- Pela inserção de símbolos especiais nas colunas.
- Pela necessidade de estandardização e correcção posterior nos valores dos dados.

Em relação às operações de estandardização e normalização, podemos conferir a necessidade em agir com vista a resolução de diversos desentendimentos nos dados, daí a necessidade da presença do esquema das fontes de dados e do esquema do DW. Geralmente, os problemas giram em torno dos conflitos estruturais dos dados. Alguns dos principais problemas de normalização e estandardização dos dados susceptíveis de melhoramento são:

- O formato ou tipo de dados não padronizados (e.g. o nome de um cliente é definido num local como *char* (30) e noutro local como *char* (45), as datas no formato YY/MM/DD e no formato DD/MM/YY).
- A estandardização de conceitos homónimos e sinónimos. O primeiro caso resulta quando o mesmo nome é usado para conceitos diferentes (e.g. o equipamento num local pode ser considerado os móveis existentes, enquanto noutro local pode ser o *hardware*). Os valores sinónimos resultam quando o mesmo conceito é descrito por vários nomes (e.g. agricultor, agricultora, camponês, camponesa).

- A imposição de regras de negócio sobre os valores dos dados em vista a obtenção de dados normalizados (e.g. nº de contribuinte – 999 999 999).
- A necessidade de mecanismos de conversão a empregar sobre os dados (e.g. sistemas métricos distintos, o peso em libras ou em quilogramas).

No que respeita às operações de correcção dos dados, devem ser tomadas posições claras sobre os dados a corrigir que se prendam além da padronização e consistência de valores. Assim, devem ser rectificados:

- Os valores com ruído (e.g. o peso de uma pessoa é 560 kg).
- O conflito no domínio dos dados (e.g. código postal).
- Os valores que não respeitam as regras do negócio (e.g. o cliente prestigio sem uma extensão de crédito associada).

A resolução da ausência de valores merece especial cuidado e deve ser abordada neste conjunto de operações correctivas. Esta situação tende a ser mais melindrosa na medida em que é preciso definir como deve ser efectuada a correcção e o preenchimento de valores por defeito [Kimball et al., 1998] [Raisinghani, 1999]. A ausência de valores nos dados pode ficar dever-se:

- Aos dados não se encontrarem sempre disponíveis.
- Ao mau funcionamento do equipamento.
- À inconsistência com outros dados armazenados e consequente supressão.
- À não entrada de dados devido a enganos ou de determinados dados não serem considerados importantes no momento da introdução dos valores.

As tarefas de integração dos dados visam a conciliação entre os dados provenientes de fontes dispersas e de natureza interna ou externa à organização. As operações mais comuns são:

- A combinação de fontes de dados, através do confronto entre o valor das chaves, ou confrontos profundos entre os atributos não chave.
- A identificação e reunião de valores duplicados (e.g. a mesma entidade encontra-se registada duplamente).
- A resolução de conflitos entre relações de dependência (e.g. uma relação num local pode ser 1:1 e noutro local essa relação pode ser 1:n).

– A criação de chaves de substituição para as diferentes linhas das dimensões, substituindo assim, as chaves das fontes. Deste modo, é possível uma melhor manutenção da integridade referencial entre as dimensões e as tabelas de factos.

Por último, as operações relativas à remoção de dados centram-se em retirar os dados considerados inválidos para o DW. Esta opção mais radical justifica-se pelo facto dos dados não serem processados convenientemente pelas operações anteriores e por isso, caso sejam carregados corrompem a qualidade dos dados e poluem o DW. Esta posição pode ficar a dever-se à utilização de uma coluna usada de forma livre e que seja impossível de compreender e corrigir, à rejeição de uma coluna sem interesse, ou à supressão de uma coluna que apresente um elevado grau de sujidade dos valores presentes e impeditivos de atitudes correctivas. Ou ainda pela natural evolução dos dados, os dados recentes e veteranos registados numa mesma coluna tendem a não se referirem à mesma característica, ou seja, a mesma coluna pode ser referida para propósitos distintos. A remoção de dados pode ainda resultar da detecção e reunião de linhas de valores duplicadas. Em qualquer circunstância, o administrador do DW, pelo conhecimento cabal do SO, deve determinar quais os dados que interessa manter, corrigir e eliminar. Assim, esta situação introduz a necessidade de uma ponderação sobre as consequências da perda de dados. A primeira respeita a averiguar se os prejuízos provocados pela ausência de dados são piores do que aqueles que resultariam da existência de dados incorrectos. A segunda resulta na possibilidade da rejeição de determinados dados provocar a necessária rejeição de outros dados de modo a manter as questões estruturais dos dados intactas, tanto ao nível da tabela como de outras tabelas (e.g. a quebra de integridade referencial) [Olson, 2003].

## Preparação dos dados

Relativamente às operações de preparação dos dados para os utilizadores finais, é possível a obtenção e cálculo de novos valores pela combinação de dados. Estas acções derivam da apresentação das características mais adequadas que os dados devem possuir, tendo em vista o interesse do negócio, a obtenção de dados valorizados e agradavelmente acessíveis aos consumidores finais. Algumas dessas operações são:

– A derivação de valores das colunas (e.g. valor_vendas_líquido + IVA $\rightarrow$ total_venda).

– A agregação de valores colunas (e.g. vendas_por_dia $\rightarrow$ vendas_por_mês).

– Os cálculos numéricos comuns (e.g. contagens e somatórios).

– A combinação e separação de atributos (e.g. 2004, 05, 31 $\rightarrow$ 31.05.2004).

– Os cálculos sobre categorizações de atributos ou construção de hierarquias conceptuais (e.g. contagens por concelho/distrito/região).

### 3.5.3 Qualidade dos dados no DW


**Figura 3-5 – Problemas nos dados no DW.**

Apesar de identificados e aparentemente resolvidos os aspectos essenciais relacionados com a qualidade dos dados durante a fase imediatamente precedente ao carregamento dos dados no DW, outras situações permanecem por resolver e só mais adiante na arquitectura dum SDW são detectáveis. Em face dos dados fluírem entre diferentes componentes e apresentarem um carácter dinâmico e evolutivo no seu conteúdo e formato ao longo do tempo, revela-se impossível a manutenção da qualidade dos dados como uma actividade de natureza estática, que uma vez conquistada para sempre estará resolvida. Na verdade, uma vez que as organizações mantêm dados provenientes de uma variedade de fontes e referentes a longos períodos de tempo, a qualidade dos dados degrada-se com o tempo (figura 3-5). Assim, uma vigilância constante sobre os dados mostra-se decisiva e oportuna. Esta é a justificação para o aparecimento das ferramentas de monitorização dos dados [Dataflux, 2004].

Neste sentido, o terceiro nível num SDW onde a qualidade dos dados pode ser questionada é o próprio repositório de dados do DW. Segundo *Inmon*, existem duas ordens de factores para que tal aconteça [18]. Em primeiro lugar, os dados residentes no DW assentam em dados históricos, muitos deles possuem significados diferentes daqueles que possuíam anteriormente, na altura da introdução dos dados, ou seja, os dados recentes e veteranos que são registados num mesmo campo tendem a não se referirem à mesma característica. Esta situação pode ser observada com recurso a varrimentos dos dados constantes no DW e que se tornem reveladoras da forma como o significado dos dados tem variado ao longo do tempo (e.g. pode ser observável que antes duma determinada data a representação do sexo de um aluno era ‘M’ ou ‘F’ e que após essa data o sexo passou a ser representado por ‘1’ e ‘0’). Assim, é possível aferir, através de varrimentos, sobre a invalidade ou deficiências dos dados perturbadoras no sistema. Em segundo lugar, a observação dos dados no DW permite analisar o estado geral dos dados, isto é, validar os domínios, os intervalos e as verificações dos cálculos sobre os dados. Estes dados vincados por um peso histórico que os descaracterizam, não são, geralmente, alvo de interrogações, e quando aliados a outros

que também não são acedidos pela irrelevância que mostram junto dos decisores, constituem um bloco de dados inertes ou monos, comummente designados por dados dormentes [Inmon et al., 1996]. As consequências da presença de dados dormentes no DW são nefastas em termos de desempenho no acesso aos dados e possíveis inconsistências entre os dados. A monitorização dos dados visa o melhoramento contínuo sobre a sua qualidade. A este propósito em [Dataflux, 2004], refere-se que a melhoria da qualidade dos dados não se trata de um projecto de uma só realização. Trata-se de um processo contínuo em vista um progressivo melhoramento dos dados em que a melhoria da qualidade dos dados não se trata somente das correcções efectuadas sobre os dados, mas também em assegurar que a organização pode continuar a usufruir de dados de elevada qualidade.

### 3.5.4 Qualidade dos dados nos DMs


**Figura 3-6 – Problemas dos dados nos DMs.**

O último local onde podem ocorrer problemas relativos à qualidade dos dados centra-se no repositório dos dados personalizados e acedidos pelos consumidores finais. Este ponto diz respeito às visões particulares, que cada consumidor ou grupo de consumidores tem sobre os dados. É o *terminus* do fluxo dos dados na arquitectura dum SDW e por isso, mostra-se como uma zona extremamente sensível, dadas as terríveis consequências que podem ocorrer ao serem difundidas informações de qualidade desajustada, quer seja em termos de interpretação, prontidão temporal, ou credibilidade das fontes (figura 3-6). Este local corresponde à última linha defensiva no que concerne à qualidade dos dados [18]. Estes dados, incumbidos de características especiais, são o resultado do esforço duma complexa estrutura inerente ao SDW e são assumidos como instrumentos de trabalho manuseados pelos diversos consumidores no exercício das suas actividades.

A constatação de resultados impuros neste momento final do processo constitutivo de informação provoca, duas consequências que impossibilitam o total agrado com o DW [18]. A primeira consequência assenta na desconfiança dos consumidores da informação sobre o sistema (e.g. um consumidor não fica agradado quando confronta os resultados obtidos através de um SDW e verifica a sua invalidade ou deficiente desempenho). Esta desconfiança inicial degenera num descrédito total sobre o sistema se o nível de qualidade não for considerado óptimo e persistir com deficiências [Strong et al., 1997]. Uma segunda consequência projecta para as causas dos problemas

verificados. Assim, a determinação do processo a montante originador dos defeitos qualitativos ou deslizes temporais responsáveis por situações impróprias, acarreta custos de auditoria, como sejam, os relativos a detecção, medição e verificação. Consequentemente, são adicionados os custos de melhoramento dos processos, pela reconversão e substituição dos processos existentes (e.g. a substituição do SGBD por estar aquém das expectativas e desempenho desejado).

De referir que o entendimento de impureza nos dados, não consiste meramente em dados errados. Pois, os dados apesar de se mostrarem correctos, os consumidores finais podem ter dificuldades no entendimento do seu significado ou no contexto da sua disponibilização. Este tipo de problemas denota a necessidade em manter metadados correctos durante todo o tempo e ao longo de todo o processo de circulação dos dados [Olson, 2003]. Os metadados referentes ao local de utilização dos dados pelos consumidores finais devem conter informações sobre como os dados se encontram codificados, como são interpretados os valores especiais, a linhagem dos dados, a data das actualizações e os níveis de qualidade dos dados disponibilizados.

## 3.6 O processo de ETL

Conforme referido anteriormente, as tarefas que precedem a divulgação dos dados aos consumidores finais são desenvolvidas na ARD e costumam designar-se por tarefas de *back-room* [Kimball & Caserta, 2004]. Estas tarefas compreendem 4 grandes áreas: extracção, limpeza, integração e carregamento dos dados no DW (figura 3-7). Estas tarefas, comummente, designadas por processo de ETL, assumem importância capital em SDWs, em especial, por serem altamente responsáveis na gestão dos dados disponibilizados. Porém, apesar da reconhecida importância, continua a ser uma área subestimada e indevidamente orçamentada em projectos de DW. A principal razão para que a área de ETL contribua com problemas adicionais em vez de apenas solucionar os existentes deve-se, essencialmente, ao desconhecimento sobre a realidade das fontes, situação que se torna explosiva no momento da transformação dos valores dos dados [Gonzales, 2004].

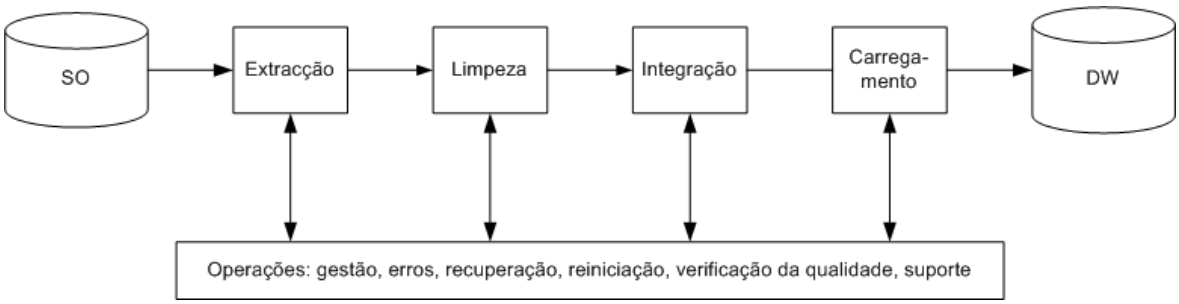



**Figura 3-7 – Fases de *back-room* dum SDW [Kimball & Caserta, 2004].**

O processo de ETL, na presente dissertação, baseia-se na adaptação da metodologia TDQM (assunto do terceiro relatório) a ambientes de DW [Wang, 1998], nomeadamente, no que concerne as actividades do processo de ETL serem encaradas como os meios de produção (semelhante a uma linha de produção de produtos comuns) responsáveis pela elaboração dos resultados (PI) divulgados aos consumidores. A exigência de um nível elevado de qualidade na realização destas tarefas é condição fundamental para a obtenção dos melhores resultados, devendo cumprir o processo de ETL a função de fiel da balança na qualidade dos dados disponibilizados pelo DW.

As características desejadas no processo de ETL devem ser: a relevância, a compreensão, o nível de detalhe adequado e a interpretação [Redman, 2004]. Tal qual os produtos produzidos, os dados de alta qualidade resultam de processos bem definidos e geridos, que criam, armazenam, movem, manipulam, processam e usam adequadamente os dados. Em [Kimball & Caserta, 2004] são tratados, especificamente, os assuntos relacionados com o conjunto das actividades de ETL e condicionam o resultado da sua acção aos requisitos e exigências impostas pela organização ou pelos meios e recursos que esta dispõe. Os requisitos podem ser encarados como constrangimentos ao sistema que tem de se adaptar para lhes dar resposta cabal, tais como: as exigências do negócio; o perfil dos dados existentes no SO; as restrições de segurança no acesso aos dados; a rapidez e frescura com que os dados têm de ser distribuídos aos consumidores finais; o armazenamento e a linhagem dos dados e a preparação dos dados de modo a serem úteis aos utilizadores. Previsivelmente, estes factores condicionam a construção do processo de ETL e podem produzir custos aterradoramente elevados no seio dos SDWs porque assimilam avultados recursos materiais, temporais e humanos.

Noutra perspectiva, podemos relevar que a qualidade dos dados num DW, em todas as suas dimensões é directamente influenciada pelas acções de ETL e consequentemente pelos requisitos a respeitar. Assim, a qualidade de conformidade da informação divulgada resulta da satisfação dos critérios previamente estabelecidos em termos das características dos dados fornecidos aos consumidores. Todavia, dadas as exigências do negócio e o insucesso demonstrado ao nível dos dados, o processo de ETL clássico mostra algumas limitações no cumprimento cabal das suas tarefas. Mesmo apesar da preocupação em centrar esforços nas tarefas comuns de ETL, os problemas ao nível dos dados continuam a abundar nos repositórios de DW e podem ser explicados por falhas contraproducentes na execução das suas actividades.

### 3.6.1 Limitações das tarefas tradicionais de ETL

O processo de ETL, além de defender o tratamento concertado dos dados recebidos das fontes, deve igualmente não ser um agente potenciador de outros problemas a jusante. Este processo constitui-se numa área susceptível de gerar potenciais problemas nos dados porque a variedade de meios envolvidos (extracção, transformação, limpeza e carregamento dos dados) pode denotar falhas, nomeadamente, quanto às especificações que devem ser respeitadas e deste modo contrariar a conformidade desejada. Em [Gonzales, 2004] é considerado que as incisivas e constantes exigências impostas por um teatro de operações, marcado pelo dinamismo, não são correspondidas pelas tecnologias tradicionais de tratamento da qualidade dos dados porque estas apresentam sérias limitações na sua acção. Cada uma das áreas de ETL pode revelar questões susceptíveis de gerarem inconformidades nos dados. Assim, no que concerne à extracção dos dados, a heterogeneidade e o estado (e.g. idade) das fontes envolvidas pode implicar uma maior ou menor complexidade no processo de extracção (e.g. ficheiros de texto, ficheiros em COBOL, símbolos especiais, imagens, mensagens de *email* e campos sobrecarregados). Em [Kimball & Caserta, 2004] são descritos 4 factores condicionantes do processo de ETL: a rapidez das acções, a postura preventiva na correcção dos dados, a identificação objectiva dos locais que degeneram a qualidade dos dados e a necessidade de perfeição no cumprimento das operações realizadas. Aliado às dificuldades de concertação de dados que não foram projectados para junções, a ausência de informações relativas às fontes (metadados) complica ainda mais este processo, situação comum em aplicações antigas porque não disponibilizam metadados e denotam especificidades muito particulares e rudimentares no modo de armazenamento [Olson, 2003] [Ballou & Tayi, 1998].

A nível da limpeza dos dados, os processos devem afiançar a obtenção de dados mais puros e geradores de valor, que sirvam de meios de suporte para a tomada de decisão. Assim, diversas operações de tratamento sobre os dados constantes nas tabelas, bem como os relacionamentos existentes entre as tabelas têm de ser realizadas [Oliveira et al., 2004]. Por isso, a compreensão integral dos dados existentes e provenientes das fontes tem de ser conquistada em vista a efectiva concretização da limpeza dos dados. No que respeita ao carregamento dos dados no DW, importa que os dados tratados se encontrem integrados e disponíveis no momento oportuno. Outras questões, designadas por pré-processamentos devem ser realizadas, como sejam: a verificação das regras de integridade; a ordenação, a agregação de valores e operações que carregam os dados nas tabelas do DW; a construção de índices que facilitem o acesso aos dados e a partição dos dados por vários locais de armazenamento [Chaudhuri & Dayal, 1997]. Deve, igualmente, ser prestada especial atenção aos dados rejeitados pelos processos antecedentes, devendo ser direccionados e armazenados num local tendo em vista a compreensão e interpretação das anoma-

lias nos valores dos dados existentes, atendendo a uma possível integração e carregamento no DW num momento posterior.

Em suma, às imperfeições constantes nos dados e que devem ser resolvidas pelo processo de ETL, não devem ser acrescidas outras, eventualmente mais graves, da responsabilidade do referido processo. Assim, a opção em incluir actividades precedentes de análise dos dados (e.g. *data profiling*), consolidações prévias dos dados e auditoria, bem como, a avaliação dos dados das fontes tem sido uma das medidas a tomar com o objectivo de obter dados mais puros pela melhor concepção das tarefas de ETL.

### 3.6.2 Políticas de circulação dos dados num SDW

É possível observar a elaboração de dados como resultado de um sistema de produção capaz de transformar dados brutos em informações úteis aos consumidores [Strong et al., 1997]. Inerente a este conceito, a política de sincronização dos dados entre os diversos patamares da arquitectura dum SDW mostra-se uma questão essencial em vista a disponibilização dos dados mais recentes aos consumidores finais em tempo oportuno e de modo consistente. A concepção arquitectural dum SDW como uma materialização de vistas em cascata, desde as fontes até aos DMs, confere a exigência em possuir os dados mais frescos a cada instante temporal [Fabret et al., 1997].

Este tema, designado por refrescamento dos dados do SDW é um assunto fulcral em relação à qualidade dos dados que o sistema disponibiliza e tem sido alvo de investigação em diversos estudos. A divulgação de dados que excedam o limite temporal previsto e aceite como plausível pelos consumidores influenciará negativamente algumas dimensões da qualidade dos dados (e.g. a oportunidade) e por conseguinte, tendem a reflectir prejuízos pela não tomada de decisão na altura certa. Do mesmo modo, a circulação inconsistente de dados por vistas materializadas desactualizadas acarreta certamente elevados custos. Estas questões revelam-se ainda mais preocupantes quando assistimos a SDWs tendentes para respostas próximas do imediatismo que o mercado exige. Logo, é necessário ter em conta as políticas de sincronização dos dados adoptadas porque a forma como se encontra implementado um SDW influência a frescura dos dados entregues aos utilizadores. Segundo o estudo realizado em [Bouzeghoub & Peralta, 2004], as tarefas de extracção dos dados das fontes para o DW, através da ARD respeitam a políticas de *push* ou *pull*. Uma política de *pull* é realizada quando a ARD executa uma consulta dos dados à fonte (e.g. a ARD executa continuamente pedidos de dados à fonte). Enquanto que perante uma política de *push*, a fonte envia os dados para a ARD (e.g. a entrega de novos dados na ARD pode ocorrer pela introdução de um *trigger* na fonte). Geralmente, os processos de propagação de actualizações recor-

rem a *triggers* nas fontes, que são activados sempre que o volume de mudanças é superior a um limite. Relativamente, à interacção existente entre o DW e os utilizadores é possível constatar, também, que os processos de consulta podem respeitar a políticas de *push* ou *pull*. Uma política *pull* permite aos utilizadores colocarem directamente as questões ao DW. Enquanto, uma política *push* possibilita aos utilizadores subscreverem determinadas consultas e o DW regularmente transmite os dados aos utilizadores. A importância das políticas a serem implementadas no SDW é um assunto fulcral, na medida que a generalidade destes sistemas se encontram associados a modos assíncronos na disponibilização dos dados aos consumidores.

A frescura dos dados nos DW é uma das traves mestras para a obtenção duma qualidade dos dados efectiva que responda às expectativas dos consumidores finais. Por isso, além das políticas de circulação dos dados pelo SDW, a frescura dos dados deve igualmente ser encarada à luz de outros factores que a influenciam e por inerência a qualidade dos dados, como sejam os perfis dos utilizadores, os perfis das fontes e os modelos de custos atribuídos. Quanto aos perfis dos utilizadores implica saber se as expectativas dos utilizadores podem ser satisfeitas e como se encontram especificados esses requisitos. Relativamente aos perfis das fontes importa a selecção de um factor de frescura das fontes, os processos de medida e os respectivos metadados que contenham a informação necessária sobre a frescura das fontes. Por último, os modelos de custos pretendem aferir sobre os custos do atraso na propagação dos dados das fontes até aos consumidores finais. Estes custos compreendem o custo de avaliação das consultas e o custo de propagação das actualizações [Bouzeghoub & Peralta, 2004].

## 3.7 Custos da qualidade dos dados

O reconhecimento da qualidade de um bem (PI) pode ser percepcionado no modo como esse bem responde aos anseios e necessidades dos consumidores (qualidade de projecto), mas também, no modo como o processo produtivo do bem (sistema de produção de informação) respeita os requisitos previamente determinados (qualidade de conformidade) [Helfert & Herrmann, 2002] [Marques, 1994]. Uma outra componente (qualidade de serviço) corresponde à capacidade de disponibilizar serviços após a entrega dos dados aos consumidores e que considerando as características destes sistemas pode ser entendida como os metadados do SDW. Estas componentes da qualidade assumem vital importância pela transversalidade verificada na obtenção de um bem capaz de corresponder ao fim destinado. Partindo do pressuposto que a informação pode ser encarada como se de um produto se tratasse, é interessante categorizar os custos associados ao processo de produção do PI de qualidade [Wang et al., 1998]. Conforme referido anteriormente, o sistema de produção do PI está predominantemente concentrado em actividades desenvolvidas na ARD.

Estas actividades são responsáveis por mais de dois terços de todos os recursos dispendidos em SDWs e a importância que ocupam é alvo de merecedora atenção porque têm acção directa na qualidade do bem apresentado para auxiliar o consumidor na tomada de decisão.

Assim, podemos encarar os custos inerentes à qualidade de conformidade em duas vertentes: os custos directos e os indirectos [Moreira, 2001]. Os primeiros respeitam aos custos de prevenção, que se encontram associados às actividades compostas pelos processos de melhoramento da qualidade e prevenção de defeitos, isto é, são os custos suportados pelas organizações de modo a implementar uma arquitectura tecnologicamente capaz da detecção e correcção dos erros que afectam a qualidade dos dados (e.g. programas de melhoramento contínuo dos dados, *software* de qualidade dos dados, formação, meios de inspecção, recolha e análise dos dados [Marques, 1994] [Moreira, 2001]. Ainda nesta categoria, podemos observar os custos de detecção ou avaliação de erros a nível da qualidade nos dados (e.g. inspecção e teste na produção, avaliação das fontes de dados, auditoria, custos de mão de obra, custos de perda de disponibilidade ou desempenho do sistema) [Cappiello & Francalanci, 2002]. A segunda categoria, os custos indirectos, resultam das rejeições, desclassificações ou reclamações dos consumidores e podem assumir tanto uma repercussão interna, isto é, se antes de chegar ao consumidor final são detectadas as falhas que violam as especificações predeterminadas (e.g. a apreciação dos defeitos, a inspecção a 100%, o trabalho de reparação), como uma repercussão externa, ou seja, são os consumidores finais que constatam falhas nos dados não condizentes com a qualidade desejada (e.g. refazer o trabalho após a entrega dos dados, os custos de imagem, a perda de confiança do consumidor) [Moreira, 2001] [Cappiello & Francalanci, 2002].

A combinação dos custos directos e indirectos gera os custos totais da qualidade. A evolução destas duas parcelas é no sentido de, quando uma aumenta, diminui a outra, o que implica a existência de um mínimo na curva da soma que configura a solução óptima (figura 3-8). As organizações que aplicam estes princípios à produção dos bens comuns estão dispostas a transferir os seus esforços de controlo de qualidade para a prevenção de falhas e a imprimir uma maior concentração no processo de concepção e fabrico dos bens. De igual modo, a produção da informação deve captar maiores atenções nas suas etapas de concepção e processo produtivo. A constatação que a produção de informação de elevada qualidade só é possível com matérias-primas (dados) de reconhecido valor tem apontado a necessidade em possuir meios adequados e tendencialmente configurados para a excelência a montante, de modo a obter a qualidade pretendida a jusante a custos admissíveis. Ora, o esforço empregue no alcance de custos globais mais baixos é resultante da subida dos custos de prevenção ser compensada pela redução dos custos de detecção de falhas nos dados e pela possibilidade em possuir dados bons logo à primeira [Moreira, 2001].

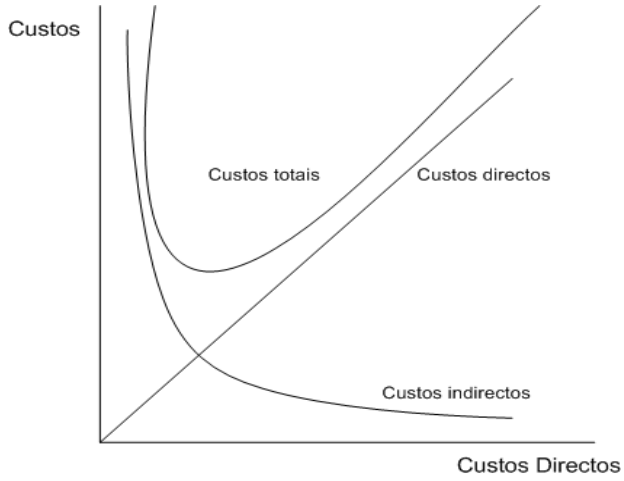



**Figura 3-8 – Custo Total da Qualidade [Marques, 1994].**

Em ambientes de DW, o SO corresponde à primeira camada da arquitectura do sistema e por isso, assumem uma importância fulcral na produção dos dados capazes de serem adequados às exigências dos consumidores. Por este motivo, não será alheia a intenção das organizações começarem a investir em meios tecnológicos capazes de garantir bons dados nas fontes. A melhoria dos dados nas fontes prevê que as iniciativas a implementar sejam coordenadas por uma equipa concentrada na garantia da qualidade dos dados e na determinação das políticas de melhoria da qualidade dos dados mais eficientes e eficazes [Kimball & Caserta, 2004]. Na realidade conforme é observável em [Kimball & Caserta, 2004], quase metade dos problemas ao nível da qualidade dos dados só pode ter um tratamento na origem, porque não existem meios tecnológicos para criar ou recriar dados não introduzidos nas fontes. É igualmente verificável que cerca de dois terços dos problemas evidenciados nos dados são melhor resolvidos nas fontes (figura 3-9).

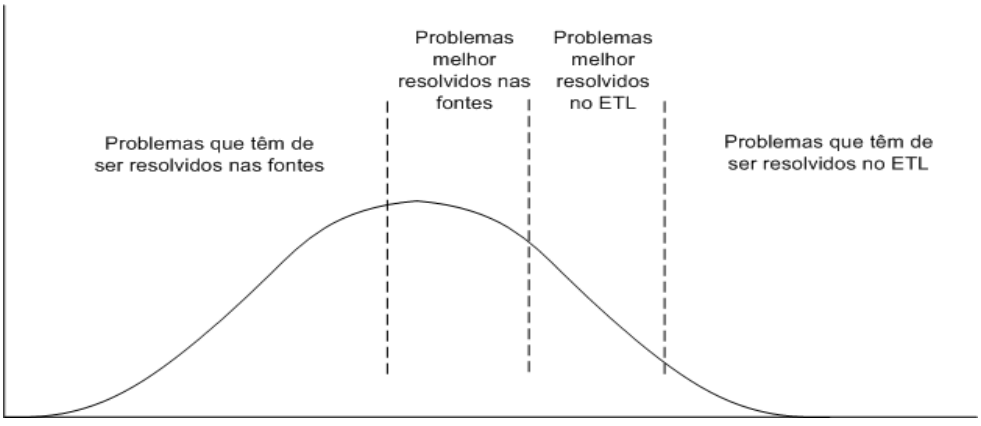



**Figura 3-9 – Políticas de qualidade dos dados [Kimball & Caserta, 2004].**

Os problemas com os dados num SDW assumem uma regra cumulativa, ou seja, os problemas que ocorrem nas camadas inferiores dum SDW mantêm-se nas camadas seguintes e geralmente tomam proporções complexas na deterioração dos dados. A introdução de valores errados nas fontes pode gerar a elaboração e publicação de informações erróneas que inviabilizem a tomada de decisões por parte dos consumidores finais e eventuais consequências no consumo de recursos necessários para a resolução dos problemas. Esta realidade constitui o designado "Principio do Funil da Qualidade", isto é, quanto mais próximo do início do processo de produção, menor será o custo da qualidade [Moreira, 2001]. Em [McKnight, 2003] é confirmada esta perspectiva, pois revela-se mais dispendioso reparar os defeitos nos dados no final da sequência estrutural do sistema do que no ponto de origem da inserção dos dados. Esta questão pode revelar-se crucial em valores dos dados cujo interesse para os ambientes operacionais se mostre reduzido e por isso, a sua recolha possa ser negligenciada. Porém, esses mesmos dados, em processamento analítico e aquisição de conhecimento, podem possuir um interesse vital para a organização. Se considerarmos que os valores desses dados são de difícil obtenção em momentos posteriores, então facilmente apercebemo-nos do aumento dos custos para a sua aquisição. Portanto, sobre este ponto podemos concluir que à medida que o PI progride ao longo do processo produtivo, mais recursos são necessários serem investidos na sua reparação (figura 3-10) [Moreira, 2001].

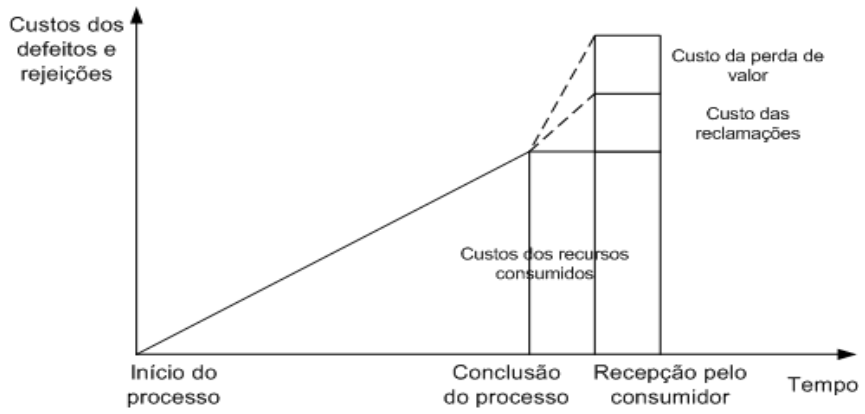



**Figura 3-10 – Princípio do funil da qualidade [Moreira, 2001].**

A tendência actual para SDWs *real-time*, caracterizados por uma janela de oportunidade reduzida, determina uma maior necessidade na antecipação da resolução dos problemas e uma atitude preventiva que evite o aparecimento de defeitos nos dados, nomeadamente, no processo de captura dos dados nas fontes. Deste modo, estaremos a contribuir para a disponibilização da informação requerida pelos consumidores no mais curto espaço de tempo, partindo do pressuposto que o processo de ETL tradicional não é suficientemente lesto e capaz de um tratamento condizente às necessidades dos consumidores e em consequência das organizações.~

# Capítulo 4

# A Gestão da Qualidade dos Dados em SDWs

A importância da qualidade dos dados no desempenho com sucesso dos SDWs conduz a exigências em termos de busca, captação e manutenção de dados de superior qualidade de modo eficaz e eficiente. Diversos estudos promovidos por académicos e profissionais reforçam esta inquietação e reconhecem este objecto de estudo como um factor crítico de sucesso na implementação dos SDWs. A natureza multidimensional e evolutiva inerente ao conceito de qualidade dos dados, induz na procura de dados caracterizados por elevado grau de correcção, oportunidade, completude, relevância e acessibilidade, deve ser preocupação cimeira dos responsáveis organizacionais. Todavia, conforme referido no primeiro capítulo, a elevada qualidade dos dados não pode ser confundida com a sua perfeição. Logo, depreende-se que a gestão da qualidade dos dados, além da natureza intrínseca dos dados, deverá essencialmente ter em linha de conta a decisão a tomar pelo consumidor final. A compreensão destas contingências permite que se encontrem reunidas as condições essenciais para a gestão ampla e objectiva das diferentes dimensões envolventes aos dados e que se revelam na relação de confiança entre os agentes decisores e um SDW.

Nos nossos dias, exige-se às organizações celeridade e elevados índices de certeza nas respostas aos mercados altamente agressivos. Os volumes de dados que circulam pelas organizações crescem exponencialmente e necessitam de ser convenientemente limpos, integrados e transformados [Gonzales, 2004]. As contingências em torno de respostas, muitas delas imediatas, implicam a frescura dos dados disponíveis para o decisor. Logo, estas questões apontam necessariamente para uma abordagem sobre toda a realidade organizacional. A implementação de plataformas tecnológicas, como os SDWs, que envolvem a globalidade da organização, não se compadecem com soluções acantonadas ou independentes no modo como a gestão dos dados é realizada.

Alguns investigadores reconhecem que, apesar de grande utilidade, as tradicionais operações e ferramentas que actuam nos SDWs (e.g. ferramentas de limpeza de dados, de controlo estatístico de processos, de monitorização dos dados) não oferecem uma aproximação sistemática na gestão da qualidade dos dados das organizações [Shankaranarayan, 2005] [Olson, 2003]. Muitas das ferramentas ou tarefas associadas não respondem convenientemente perante enormes quantidades de dados e necessidades de frescura destes. A velocidade dos processos de limpeza, transformação e absorção dos dados pelo DW mostra-se crítica para sucesso do próprio sistema e consequentemente da organização [Gonzales, 2004].

Neste contexto, o conceito de PI parece dar resposta adequada tendo em vista a conquista duma elevada qualidade nos dados fornecidos aos consumidores finais. Algumas investigações sobre a administração da qualidade dos dados [Wang et al., 1998] [Wang et al., 2003] [Wang, 1998] [Kahn et al., 2002] [Shankaranarayan, 2005] [Pierce, 2004a] [English, 1999], concluem que as organizações devem gerir os dados de modo análogo aos produtos convencionais e que os processos de elaboração e desenvolvimento inerentes devem ser compreendidos e geridos de modo a assegurar a entrega de PIs de elevada qualidade [Wang et al., 2004]. A assumpção que a qualidade dos dados deverá ser tratada de modo análogo à qualidade dos produtos convencionais enaltece a preocupação cimeira face aos dados em detrimento dos sistemas que os suportam [Pierce, 2004a]. Para entender as implicações da fraca qualidade dos dados em SDWs torna-se necessário traçar o problema da qualidade ao longo do PI e dos processos produtivos responsáveis pela sua criação. É com base nestas premissas que o presente capítulo, visa à luz do exposto em [Brackett, 1996], o conhecimento e compreensão da qualidade do repositório de dados no DW, para assim, conhecer o estado da qualidade dos dados existentes.

## 4.1 O SDW como produto-informação

A adopção de princípios e regras de gestão oriundas da produção, distribuição e qualidade dos produtos convencionais ao domínio dos dados tem-se tornado uma realidade conforme comprovam diversas investigações realizadas. A generalidade das investigações que recorrem a estes pressupostos teóricos, comprovados no terreno, defende que a gestão dos dados deve ser realizada de modo similar à gestão dos produtos comuns [English, 1999] [Shankaranarayan, 2005]. A aproximação de PI perspectiva que o resultado final a obter por um sistema de informação corresponde a um produto produzido pelo próprio sistema [Wang et al., 1998]. Para tal, o entendimento dos requisitos a respeitar na construção dum PI e o conhecimento do conjunto dos processos responsáveis para a obtenção de um produto mostra-se essencial.

Os processos de construção dos PIs correspondem ao designado sistema de produção de informação e são responsáveis por adicionar valor aos dados ao longo da cadeia produtiva. O valor atribuído a um PI determina a utilidade dos dados nas decisões a tomar pelos consumidores. O PI pode ser percepcionado como uma colecção de componentes ou elementos de dados que devidamente integrados procuram satisfazer as expectativas dos utilizadores. Em vista a obtenção deste fim, o tratamento dos PIs pelas organizações deve ser realizado, com base em quatro princípios [Wang et al., 1998]:

- Compreender as necessidades dos consumidores: as necessidades dos consumidores devem ser claramente estabelecidas e compreendidas durante todas as fases do desenvolvimento e produção dum PI.
- Gerir o processo de produção de informação: o processo deve ser bem definido e conter os controlos adequados, incluindo segurança na qualidade, inspecção, a gestão da produção e do tempo de entrega.
- Gerir o ciclo de vida dum PI: o grau e a frequência das mudanças num PI dependem do tipo e natureza da informação, das tarefas de suporte informativo e das mudanças do contexto em que a informação é usada.
- Designar um gestor para um PI: este gestor é responsável pela coordenação e gestão dos fornecimentos dos componentes para a realização dum PI, os processos de produção e a entrega dos produtos resultantes aos consumidores.

Em decorrência do exposto, tal qual o sistema de produção de bens comuns, os SDWs podem ser vistos como sistemas de produção responsáveis por converter dados em estado bruto em PIs úteis como resposta perante ambientes de decisão dinâmicos. A proposta de PI tem ganho aceitação por parte dos académicos e profissionais devido a diversos factores [Shankaranarayan, 2005]. Primeiro, a produção dum PI é realizada de modo similar a um produto convencional. As matérias-primas, os meios de produção, o processamento e a inspecção são relacionadas nas etapas do processo produtivo e os componentes podem ser processados como numa linha de produção. Segundo, diferentes PIs podem partilhar os mesmos processos ou dados de entrada, o que permite uma gestão realizada sobre grupos homogéneos (e.g. determinados grupos de dados sujeitos a um conjunto de processos de transformação iguais). Esta abordagem de agrupamento dos PIs permite efectuar uma gestão global sobre toda a classe considerada ou actuar sobre determinadas especificidades a corrigir nas classes [Shankaranarayan et al., 2003]. Por último, a possibilidade em encarar a informação divulgada aos consumidores finais de modo similar aos restantes produtos, permite a adopção de princípios comprovados nos processos produtivos convencionais, que

promovem a melhoria da especificação de requisitos, o desenvolvimento, a produção e a distribuição dos PIs. É o caso da TQM, como metodologia assente numa acumulação de investigações e aplicações práticas de âmbito multidisciplinar e composta por princípios, regras e técnicas de melhoramento contínuo dos produtos. A validade desta metodologia induziu a sua aplicação e adaptação ao domínio dos dados constituindo a TDQM, como plataforma formada por conceitos e procedimentos para a definição, medição, análise e melhoramento dos PIs [Wang, 1998].

A importância na definição de melhores PIs conduz a um aprimoramento de técnicas relativas ao levantamento dos requisitos dos utilizadores em vista atingir os seus anseios de modo mais eficaz. Esta preocupação pode ser constatada na investigação [Pierce, 2004a], que conceptualiza um PI com recurso a técnicas oriundas do *marketing* de produtos e serviços convencionais. Assim, são acrescentadas às técnicas de gestão da produção de produtos comuns, algumas técnicas que promovem o desenvolvimento e desenho dos próprios PIs, em especial, as relativas à qualidade e prioridade dos requisitos dos consumidores. A definição dos PIs deve resultar do balanceamento adequado das diversas variáveis que influenciam directamente o produto final, como sejam os requisitos dos utilizadores, os custos associados, as políticas de segurança e a distribuição das informações. O balanceamento das variáveis a ponderar na elaboração dum PI denomina-se *product-mix* e serve de base para a gama dos PIs a elaborar. O decisor é assumidamente considerado o principal interveniente com o SDW porque é o consumidor final e por isso, reconhece e valida os dados apresentados. Porém, outros intervenientes intrínsecos aos SDWs desempenham papéis de elevado destaque em vista a obtenção duma elevada qualidade dos dados. Cada um deles apresenta os seus próprios requisitos no que respeita aos dados, daí ser necessário avaliar os níveis da qualidade dos dados em todas as suas vertentes e exigências.

O paradigma PI e a metodologia TDQM justificam a utilização dos seus conceitos e princípios no decurso da investigação em curso pela aplicação ao domínio dos SDWs. Em cada etapa do processo produtivo dum SDW (extracção, limpeza, transformação, integração e carregamento) um PI é criado (*output*) e corresponderá a um *input* para a fase seguinte. Os PIs que vão sendo realizados ao longo da cadeia produtiva podem ser percepcionados como vistas materializadas. A própria natureza dos SDWs estabelece a construção de vistas sobre os dados que são materializadas e periodicamente refrescadas para um melhor e mais rápido acesso aos dados. Os gestores questionam o sistema e este responde quase sempre através desses dados previamente materializados [Bouzeghoub & Peralta, 2004]. Ressalvando o eventual objectivo específico da constituição dum SDW, o tipo de acessos apresentado e designado por consultas de tipo *ad hoc* aos dados constitui-se no acesso mais comum e no meio, potencialmente, mais interessante para a criação de vantagens competitivas geradas e alicerçadas em DWs. Assim, com base neste pressuposto,

pode-se considerar um DW como se dum PI se tratasse. Esta assumpção possibilita o tratamento esquemático do conceito, como forma de tornar mais fácil a identificação dos problemas dos dados e o entendimento das origens desses problemas. Esta alusão foi recentemente apresentada na investigação [Shankaranarayan, 2005], na busca de um melhor entendimento dos processos inerentes ao sistema de produção de dados (SDW), dos problemas a nível da qualidade dos dados e na introdução de mecanismos para medir a qualidade dos dados.

## 4.2 Propostas de melhoria da qualidade dos dados

Diversos investigadores têm fornecido propostas que visam assegurar a melhoria da qualidade dos dados existentes nos sistemas de informação em geral e nos SDWs em particular. Em face da diversidade de orientações das investigações relativas à problemática da qualidade dos dados, iremos proceder ao levantamento de um conjunto de estudos e propostas relevantes para o tema em causa e simultaneamente que se mostram complementares sobre o objecto de investigação. Deste modo, pretendemos responder perante duas vertentes: definir uma plataforma teórica com créditos firmados pela sua aplicação neste domínio e paralelamente servir de base orientadora ao presente trabalho. Os traços de complementaridade que algumas investigações revelam, possibilita a concertação das suas acções em vista a disponibilização de dados de superior qualidade, utilizáveis e que acrescentem valor aos dados primários. Na generalidade, as propostas apresentadas socorrem-se de estudos anteriormente divulgados para estabelecer um alicerce teórico que sustente as investigações a apresentar. Assim, muitas das propostas que iremos apresentar não se afirmam como um fim em si mesmo, porque exigem a compreensão e apreensão de outros estudos. O recurso a investigações anteriormente enunciadas revela-se igualmente interessante na medida que permite validar os modelos, técnicas e conceitos teóricos adoptados.

### 4.2.1 A proposta de Redman

Em [Redman, 1995] são descritas três estratégias em vista o melhoramento da qualidade dos dados: a identificação do problema, o tratamento dos dados como assunto e a implementação de sistemas de qualidade mais avançados. A estratégia inicial passa pela identificação das irregularidades nos dados. Este aspecto é geralmente negligenciado pelas organizações, que não efectuam medições sobre os dados e consequentemente se confrontam com dificuldades na detecção de eventuais imperfeições destes.

A implementação da segunda estratégia, o tratamento dos dados como assunto, consiste na adopção dos dados como uma área de gestão autónoma que desempenha um papel crucial nas

organizações [Redman, 1995]. Algumas das actividades de gestão dos dados envolvem a sua inventariação, o reconhecimento do valor dos seus processos criadores, a atribuição de responsabilidades pela qualidade dos dados e o estabelecimento dum relacionamento do tipo fornecedor/cliente dos dados [Leitheiser, 2001].

Por fim, a estratégia de implementação de sistemas de qualidade mais avançados requer a definição de processos para a detecção e correcção de erros verificados; a implementação de um processo de gestão para a descoberta e eliminação das causas geradoras das deformidades nos dados e o redesenho dos processos de modo a serem menos susceptíveis de produzirem defeitos nos dados [Leitheiser, 2001].

### 4.2.2 Quality Function Deployment

O *Quality Function Deployment* (QFD) é um método para o planeamento da qualidade que tem os requisitos dos consumidores como componente fulcral, foi introduzido no Japão em 1966 e tem sido adaptável ao domínio da qualidade dos dados partindo da premissa do conceito PI [Helfert & Radon, 2000]. A consideração do método ao domínio da qualidade dos dados tem sido realizada desde meados dos anos noventa em algumas investigações [Redman, 1996] [Helfert & Radon, 2000] [Vassiliadis, 2000]. Este método baseado, nos requisitos dos consumidores, desenvolve pontos-chave referentes à garantia da qualidade ao longo do processo produtivo. O método exige o permanente envolvimento das diferentes equipas de trabalho ao longo de quatro fases:

- Os critérios de qualidade respeitados atendendo aos requisitos dos consumidores.
- A disposição das características em cada uma das componentes do produto.
- A definição das especificações do processo de produção através das especificações das componentes do produto.
- A definição das ferramentas para a produção e qualidade de inspecção de modo a assegurar as exigências no processo produtivo.

Este método pode ser aplicado ao planeamento e medição da qualidade dos dados em SDWs, pela adaptação das etapas ao domínio do DW [Helfert & Radon, 2000]. A figura 4-1 ilustra este método que permite o planeamento da qualidade nos dados. O elemento central deste método é designado por *casa da qualidade* e consiste no documento principal resultante das quatro fases consideradas. Primeiro, são modeladas as necessidades e expectativas dos consumidores (requisitos objectivos e subjectivos) e é produzida uma lista dos objectivos a atingir, comummente designada de *WHATs*. Segundo, são modeladas as soluções técnicas, ou seja, as características da

informação tendo em vista responder aos requisitos avançados no ponto anterior (*HOWs*). Terceiro, é efectuado o cruzamento entre os requisitos dos utilizadores e as características da informação, sendo preenchida o interior da casa, que se designa por *matriz de relacionamentos*. Por último, são identificadas as relações entre os factores técnicos, o que exige a negociação entre os intervenientes do DW sobre pontos de interesse antagónicos. Esta negociação corresponde ao telhado da casa e designa-se por *matriz de correlação* [Helfert & Radon, 2000] [Vassiliadis, 2000]. Em seguida, devem ser construídas as avaliações competitivas, que correspondam à comparação entre os produtos competitivos e os produtos da organização. A avaliação é separada em duas categorias: a avaliação dos consumidores (requisitos dos utilizadores) e a avaliação técnica (características da informação). Esta avaliação permite validar a *matriz de correlação* e derivar valores passíveis de ser alcançados nas características da informação. Consequentemente, são concedidas as prioridades aos requisitos dos consumidores, que correspondem a um bloco de colunas por cada requerimento dos consumidores e contêm o grau de importância e o objectivo a alcançar. Por último, são definidas as prioridades técnicas das características da informação, através da anotação do grau de dificuldade, o objectivo a alcançar, e os pesos relativos e absolutos [Helfert & Radon, 2000] [Vassiliadis, 2000].

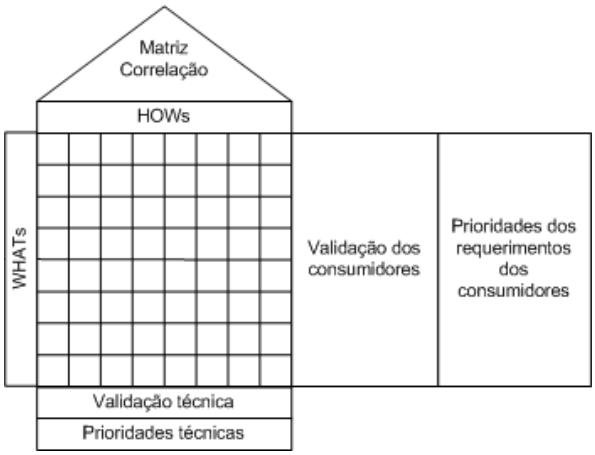



**Figura 4-1 – Casa da Qualidade [Vassiliadis, 2000].**

### 4.2.3 Total Data Quality Management

A metodologia TDQM [Wang, 1998] segue a aproximação TQM, adaptando-a para a avaliação e melhoria da qualidade dos dados na generalidade dos sistemas de informação. A implementação da metodologia pressupõe a compreensão do conceito PI, como informação resultante das matérias-primas (dados em bruto) operadas por um sistema de informação. O PI é analogamente comparado a um produto comum, em especial, no que respeita às propriedades que o caracterizam e

permitem aferir da sua qualidade. O ciclo característico da metodologia TQM: *Plan, Do, Check, Act* (PDCA) é ajustado e dá procedência ao ciclo composto por: definir, medir, analisar e melhorar. A primeira etapa do ciclo consiste na definição das dimensões ou características da qualidade dos dados em termos da funcionalidade que devem garantir para os consumidores dos dados. Significa igualmente definir os requisitos dos PIs por parte das diferentes ópticas dos intervenientes (produtores, administradores e consumidores). Estabelecidos os requisitos dos intervenientes e as características dos dados, é possível definir o sistema de produção da informação. A etapa de medição visa a produção de métricas para a avaliação da qualidade dos dados e do sistema de produção de dados. A etapa seguinte pretende a identificação das causas dos problemas na qualidade dos dados. Nesta etapa, recorre-se a métodos específicos capazes de objectivamente produzirem factos sobre defeitos nos dados (e.g. controlo estatístico de processos). Simultaneamente, é calculado o impacto da fraca qualidade dos dados no seio da organização. Por último, a etapa de melhoramento corresponde às iniciativas visando a eliminação das causas dos problemas e consequente melhoria da qualidade dos dados. Algumas iniciativas passam pelo realinhamento das características chave da informação e as necessidades da organização ou com o processo de produção de informação [Wang, 1998] [Vassiliadis et al., 1999].

A metodologia compreende a constituição dum PI em torno de três vectores fundamentais: as características intrínsecas, a qualidade associada e o sistema de produção responsável pela sua elaboração. A figura 4-2 mostra o carácter dinâmico e iterativo na implementação da metodologia, ou seja, a definição dum PI como resposta a uma exigência, num dado momento, pode não responder às necessidades num momento posterior [Wang, 1998].

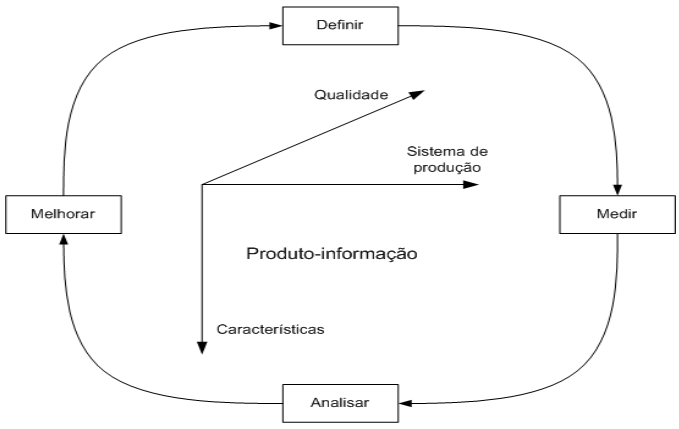



**Figura 4-2 – Esquema da metodologia TDQM [Wang, 1998].**

A aplicação da metodologia exige da organização a capacidade em [Wang, 1998] [Pierce, 2004a]:

- Relacionar os PIs e a sua qualidade associada em termos de negócio.
- Estabelecer uma equipa consistente, responsável e chefiada por um gestor, para o desenvolvimento dos PIs.
- Formar os intervenientes com os PIs para as actividades de gestão e validação (e.g. a definição dos critérios e das métricas de validação e a análise das causas responsáveis pela fraca qualidade dos dados).
- Institucionalizar a melhoria contínua dos PIs.

### 4.2.4 Information Product Map

Para avaliar as implicações da fraca qualidade dos dados é necessário entender as etapas do processo produtivo da informação, bem como o impacto provocado por tempos de espera associados a essas etapas [Shankaranarayan et al., 2003]. A proposta *Information Product Map* (IPMAP) possibilita uma compreensão intuitiva e a representação visual do processo produtivo dum PI [Shankaranarayan et al., 2000]. Esta abordagem consiste num método completo para a representação e melhoramento dos dados disponibilizados aos diversos intervenientes no processo produtivo, através da detecção das actividades que contribuem para diminuir o nível da qualidade dos dados no processo produtivo ou que são responsáveis pela degeneração da linhagem destes. Tendo em vista a dar resposta a estas preocupações, a proposta visa alcançar os seguintes objectivos [Scannapieco et al., 2003]:

- Visualizar as fases críticas do processo produtivo que afectam a qualidade dos dados.
- Visualizar os fluxos de dados em todo o processo, para avaliar se os tempos de espera são os estimados para a entrega dos PIs.
- Medir o nível de qualidade dos dados nas diferentes etapas do processo produtivo.
- Melhorar continuamente o processo produtivo.

O IPMAP permite ao decisor a visualização da distribuição dos dados e outros recursos ao longo do processo produtivo de criação dum PI. A combinação do IPMAP com um adequado repositório de metadados e as capacidades proporcionadas pela aplicação da TDQM, permitem ao decisor, em todas as etapas, a disponibilização de um conjunto de informações relativas ao nível de qualidade dos dados assegurado pelo sistema, como sejam: a identificação do processo envolvido, a localização física, o sistema usado, a composição do produto ou subproduto e a organização envolvida na criação dum PI. O IPMAP é uma extensão do sistema de produção proposto em [Ballou et al., 1998]. Enquanto, este último visa somente a obtenção da qualidade no produto final, o

IPMAP acrescenta a compreensão e representação do processo de produção do PI [Shankaranarayan et al., 2003]. O IPMAP compreende cinco etapas [Scannapieco et al., 2003]:

- A catalogação dos PIs: para inventariar os PIs e as características que os individualizam (e.g. a natureza, os consumidores e os processos envolvidos).
- A identificação dos PIs críticos: no sentido de enveredar esforços para uma melhoria da qualidade (e.g. os responsáveis por gerar perdas avultadas por falta de qualidade).
- A definição dos requisitos de qualidade para os PIs críticos: para conhecer os requisitos considerados como um produto de qualidade. A qualidade na análise dum PI determina a necessidade em constituir métricas sobre o produto ou os componentes que o constituem.
- A construção do IPMAP e do repositório de metadados: o IPMAP irá descrever graficamente, com recurso a oito blocos de construção, o processo de elaboração dum PI. Cada bloco é identificado por um nome único e descrito por um conjunto de atributos, que podem corresponder aos metadados.
- A avaliação e o melhoramento da qualidade dum PI: uma vez construído o IPMAP, podem-se implementar medidas visando a melhoria da qualidade dum PI. Esta etapa procura prevenir, detectar ou corrigir algumas anomalias nos dados. Porém, as inspecções não detectam todos os tipos de erros, incumbido à organização (ou ao administrador dos dados) a tarefa de ponderar o peso dos erros detectados e verificar se cumprem os compromissos estabelecidos. Na avaliação, é possível recorrer a matrizes de controlo, que são capazes de relacionar os problemas dos dados aos controlos de qualidade, para detectar e corrigir os problemas dos dados ao longo do processo produtivo [Pierce, 2004a].

Logo, pode-se referir que o propósito do IPMAP consiste em modelar todo o processo produtivo e compreender o modo como os vários componentes dum PI agem em conjunto. Em [Shankaranarayan, 2005] refere-se que o IPMAP oferece três capacidades de gestão da qualidade dos dados e de implementação da TDQM: a estimativa do tempo de entrega, o alcance e o rastreio. O tempo de entrega dum PI corresponde ao tempo dispendido na elaboração da informação ou de um componente. O tempo necessário para a execução dum PI condiciona os decisores a ponderarem sobre PIs alternativos que respeitem os critérios predefinidos em tempo considerado aceitável. Esta questão revela-se particularmente interessante pela necessidade de um elevado grau de frescura dos dados divulgados por um SDW. A capacidade de alcance consiste na identificação de todas as etapas constituintes dum PI passíveis de visualização a partir de uma etapa descrita no IPMAP. A importância desta capacidade resulta, especialmente, da identificação do impacto provocado por erros ao nível da qualidade. Se uma unidade de dados contida numa etapa do IPMAP

denotar falhas de qualidade, então afectará todas as etapas do processo produtivo que se encontram a jusante do local considerado. A capacidade de rastreio consiste em identificar ou traçar a sequência de uma ou mais etapas que precedem uma qualquer etapa. Assim, é facultada a visualização da árvore geneológica dos dados e permite ao administrador dos dados ou ao decisor a capacidade de averiguação sobre as causas das quebras de qualidade dos dados nos PIs. Seguidamente, iremos listar e descrever os oito símbolos de construção do IPMAP (tabela 4-1). Em IPMAP, cada símbolo é descrito por um conjunto de atributos.

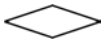
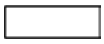
| Símbolo | Designação | Descrição |
|---|---|---|
|  | Fonte de dados (matéria-prima) | Representa cada fonte de dados necessária para a produção dum PI. Associado a este bloco encontra-se a unidade de negócio responsável pela fonte de dados, o processo usado para captura dos dados e o sistema que armazena esses dados. |
|  | Consumidor (output) | Representa o consumidor dum PI. O consumidor especifica os elementos de dados que constituem um PI, implicando a necessidade deste ser identificado antecipadamente. Associado a este bloco encontra-se a organização/unidade de negócio/departamento encarregue do PI, o nome da entidade que irá usar o produto e o conjunto de dados constituintes do PI. |
|  | Qualidade dos dados | Indica a verificação da qualidade dos dados que compõe um PI. A avaliação dos componentes possibilita a produção de informação livre de erros. Associado a este bloco existe uma lista de verificações de qualidade dos dados executadas em cada um dos componentes. Os *inputs* neste bloco são as fontes de dados e alguns componentes de dados (e.g. verificar domínios, verificar ausência de valores e autorizações). |
|  | Processamento | Interpreta as manipulações, os cálculos ou combinações que envolvam, parcial ou totalmente, os dados provenientes das fontes ou dos componentes para a obtenção dum PI. Quando este bloco é usado com o propósito específico de limpeza ou correcção dos dados introduzidos, então passa a ser designado como bloco de correcção dos dados. |
|  | Armazenamento dos dados | Este bloco é usado para indicar a captura de elementos de dados em bases de dados ou ficheiros para futuras utilizações. Estes blocos podem ser usados para representar os elementos de dados (matérias-primas ou componentes) que esperam processamento futuro ou são capturados como parte do inventário de dados na organização. |
|  | Limites do negócio | Identifica as matérias-primas ou componentes dos dados que são transmitidas para outra organização/unidade de negócio/departamento. O papel deste bloco consiste em realçar problemas de qualidade dos dados que podem aparecer pelo cruzamento entre organizações ou unidades de negócio. |
|  | Decisão | Em sistemas de produção de informação mais complexos pode ser necessário direccionar condicionalmente os elementos de dados para conjuntos de blocos para processamento futuro. Nestes casos, um bloco de decisão é usado para representar as diferentes condições a avaliar e os correspondentes procedimentos que irão acolher os dados provenientes dessa avaliação (e.g. os dados relacionados com o nascimento podem ser usados para gerar um certificado de nascimento ou relatório sobre estatísticas de nascimento). Cada objecto representa um PI e pode usar os mesmos dados (componentes e matérias-primas) na sua produção. |
|  | Limites do sistema de informação | Indica o reflexo entre as mudanças das matérias-primas ou elementos componentes dos dados no movimento de um sistema de informação para outro sistema de informação, especificando assim, os sistemas envolvidos. As mudanças podem ser interiores ou exteriores às unidades do negócio. As matérias-primas ou componentes podem circular por limites do negócio ou entre sistemas de informação (e.g. movimentação de elementos de dados de um SGBD para outro). |

**Tabela 4-1 – Componentes do IPMAP [Scannapieco et al., 2003] [Shankaranarayan et al., 2003].**

### 4.2.5 A proposta de Shankaranarayan

Recentemente, uma nova proposta [Shankaranarayan, 2005], baseada nas duas investigações precedentes, tem como objectivo a gestão da qualidade dos dados em SDWs. O estudo propõe a transferência dos princípios e conceitos da TDQM ao domínio dos SDWs, através da aplicação da aproximação PI e da representação gráfica IPMAP. Neste contexto, um SDW é visto como um sistema de produção composto por diferentes processos capazes de operar e integrar os dados provenientes de diversas fontes e de produzir dados relevantes na condução das actividades de negócio. A compreensão sobre a qualidade dos dados num SDW determina o conhecimento cabal sobre os processos envolvidos e o modo como o grau de qualidade dos dados em cada etapa do processamento dos dados é influenciado pelas etapas precedentes. A aferição sobre a qualidade dos dados nas diferentes etapas de processamento dos dados prevê a disponibilização de meta-dados associados em cada etapa ou patamar do processo produtivo [Shankaranarayan, 2005].

O esquema de modelação do IPMAP permite a representação das diferentes etapas de produção dos PIs ao longo dum SDW. Recorreremos a um exemplo ilustrativo da aplicação deste modelo em SDWs, especialmente, pela visualização da construção dum DW como se de um produto se tratasse. O DW considerado no exemplo é composto por duas tabelas dimensão e uma tabela de factos. Uma sequência de etapas representativas para a construção do DW é mostrada na figura 4-3. Os dados das fontes (FD1 e FD2) são extraídos em função da contextualização de cada decisão a tomar pelos utilizadores. Os dados extraídos pelos processos (PE1 e PE2) são sujeitos a operações de limpeza pelos processos (PL1 e PL2). Os dados limpos oriundos da FD1 são corrigidos pelo processo (PT1). Em seguida, estes dados são combinados com os dados da FD2 através do processo de integração (PI1), inspeccionados pelo processo (I1) e armazenados em (A1). As restantes tabelas (dimensões e factos) podem igualmente ser representadas de modo análogo conforme ilustra a figura 4-4. O armazenamento (A3) pode ser combinado a outros armazenamentos temporários (A2 e A3) através de um processo de transformação (PT) e o resultado carregado pelo processo (PC1) no DW (A4) (figura 4-5).

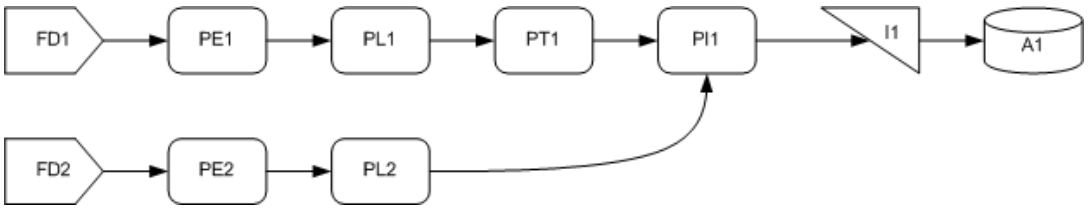



**Figura 4-3 – IPMAP do armazenamento dos dados referentes a tabela de factos na ARD.**

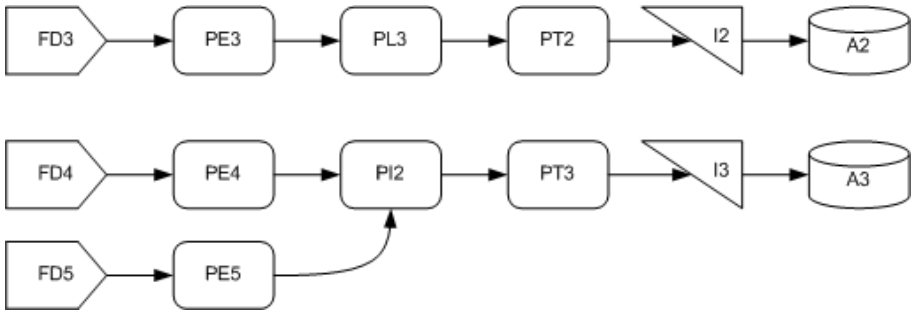



**Figura 4-4 – IPMAP do armazenamento dos dados das tabelas dimensão na ARD.**

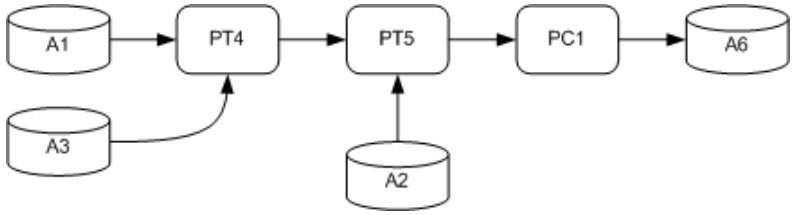



**Figura 4-5 – IPMAP que representa a transformação e carregamento dos dados no DW.**

Este modelo compõe-se por quatro estádios. No primeiro estádio é prevista a definição dos requisitos dos metadados para a gestão da qualidade dos dados. A garantia de uma boa qualidade dos dados deve prever atributos de metadados que contenham a documentação de processos e procedimentos, como sejam sobre a captura dos dados, o armazenamento, as regras de transformação, as métricas e referências ao uso dos dados. O segundo estádio, descreve a representação do IPMAP para a captura dos metadados e o modo como estes são comunicados ao decisor. Deste modo, auxilia-se o decisor sobre qual o resultado em cada momento processual do SDW, como é obtido esse resultado, onde é a localização física, quem é o responsável e quando é executado o processamento. O terceiro estádio determina a associação entre o IPMAP e os metadados do DW. Finalmente, é efectuada a avaliação da qualidade dos dados dentro do conceito IPMAP [Shankaranarayan, 2005].

### 4.2.6 Total Information Quality Management

Em [English, 1999] é apresentada uma metodologia para a qualidade da informação que tem a sua origem na TQM e nos catorze pontos para a qualidade enunciados por *Edwards Deming*. A proposta *Total Information Quality Management* (TIQM) consiste em transpor a metodologia, nos seus princípios, métodos e técnicas, para o domínio da informação. O objectivo geral a atingir consiste em institucionalizar uma atitude de melhoria contínua dos processos envolventes aos dados, por oposição a atitudes de rectificação pontual dos defeitos nos dados. Em vista a maximização dos recursos envolvidos no processo de limpeza, as organizações devem assumir a correc-

ção dos problemas nos dados como uma etapa única e complementar esta iniciativa, com a melhoria dos processos causadores dos defeitos nos dados [English, 2003b]. Importa salientar o facto da semelhança entre esta proposta e a divulgada em [Wang, 1998]. Ambas assentam na analogia entre a qualidade de um produto comum e a qualidade da informação produzida por um sistema de informação. Adoptam, igualmente, os princípios estabelecidos e consensualmente reconhecidos da TQM. No contexto desta dissertação, interessa salvaguardar o interesse e aplicação da metodologia e não em contribuir sobre a originalidade na adopção do conceito.

A TIQM compõe-se por seis processos que prevêem a melhoria contínua da informação, conforme mostra a figura 4-6. Os três primeiros respeitam a processos de avaliação. O ciclo inicia-se com a avaliação dos metadados, o que inclui o acesso à definição dos dados e à sua arquitectura. O segundo processo prevê a avaliação da qualidade da informação, através da definição do processo de medição da informação com vista alcançar as características estabelecidas. O terceiro processo efectua uma análise dos custos imputáveis à fraca qualidade das informações. É realizada, igualmente, uma análise de custo e benefício das acções de melhoria a implementar. Em seguida, são realizados os projectos de correcção, de transformação e de migração dos dados (e.g. dados em trânsito para o DW). O quinto processo respeita à melhoria dos processos de produção/consumo da informação. Este processo recorre à implementação do ciclo de *Shewhart* – PDCA. Por último, o sexto processo estabelece o ambiente organizacional para a qualidade da informação, ou seja, consiste num processo de promoção de mudança da cultura organizacional, que fomente a valorização dos consumidores de informação, a excelência dos processos e o hábito da melhoria contínua dos processos [English, 2003b] [Amaral et al., 2002].

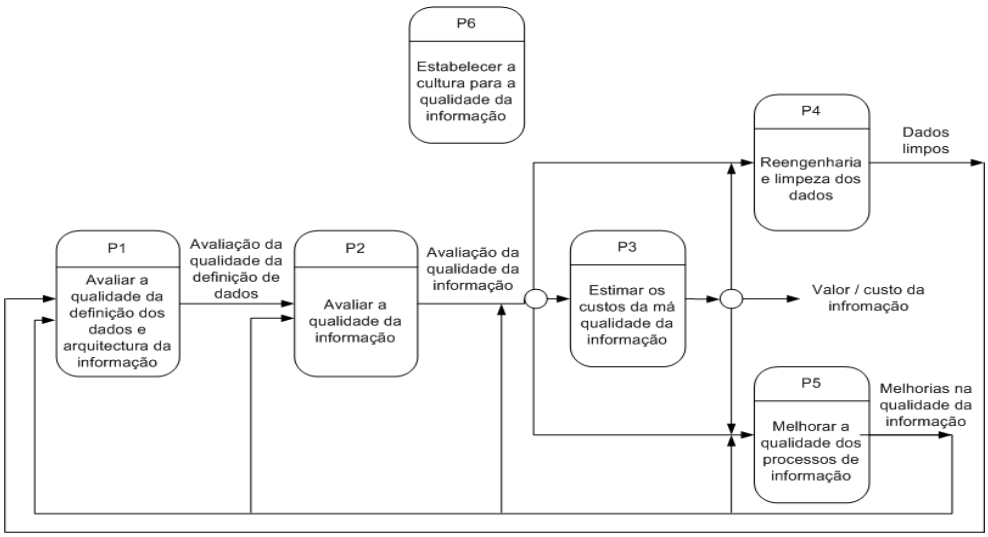



**Figura 4-6 – Processos da TIQM [Amaral et al., 2002].**

A aplicação desta metodologia não determina a sua adopção na totalidade. É possível tratar uma questão relativa à qualidade dos dados que recorra apenas a uma parte dos processos e o alheamento relativamente aos restantes. Esta versatilidade pode ser constatada em [Amaral et al., 2002], que descreve o processo de migração ocorrido entre duas plataformas informáticas. É igualmente importante salientar que os seis processos apresentados correspondem a um patamar inferior de especificação da metodologia. Em cada um dos processos é possível obter-se um maior nível de detalhe que discrimine os sub-processos a realizar e que configure uma estrutura de superior pormenorização.

### 4.2.7 A proposta de Olson

A proposta de Olson parte do princípio que o resultado das investigações sobre a qualidade dos dados conduz a factos que uma vez agrupados formam assuntos sobre os dados. Os assuntos correspondem a problemas ao nível da qualidade dos dados e representam a etapa primeira na resolução e melhoria da qualidade dos dados. Assim, em vista o aperfeiçoamento dos dados é apresentada uma sequência de actividades a realizar. A primeira consiste na conversão dos factos em assuntos. O recurso a técnicas de análise de dados produzem factos relacionados com as deformidades nos dados que podem indiciar a existência de valores errados. O agrupamento dos factos identificados leva a obtenção de um conjunto de assuntos que definem problemas a resolver. O recurso a métricas como medidas de avaliação sobre os níveis de qualidade verificados pode revelar a ocorrência de factos (incumprimentos) sobre os dados. Uma simples estatística pode resultar num assunto (e.g. a detecção que 20% das linhas da tabela dos alunos não contêm a identificação da localidade).

A segunda actividade refere-se à avaliação do impacto. Assim, cada assunto abordado necessita ser observado em termos do potencial impacto na organização. É essencial a justificação dos esforços no desenvolvimento de acções correctivas ou conhecer o valor de retorno para a organização. Conforme referido anteriormente, os custos resultantes das falhas de qualidade registadas nos dados são de muito difícil determinação e geralmente a previsão dos benefícios a atingir revela uma dificuldade ainda mais acrescida. Porém, a análise financeira provocada pelo melhoramento dos dados deve ser avaliada porque permite a confrontação entre os custos e os benefícios resultantes do investimento. Idealmente, os investimentos a realizar devem resultar da escolha da opção que mais se adequa às necessidades da organização.

A actividade seguinte incide sobre a investigação das causas dos problemas nos dados. Nesta fase, procura-se descobrir as causas dos factos ou das anomalias verificadas. A investigação das

causas nem sempre pode ser concretizada (e.g. inclusão de fontes de dados externas num SDW). *Olson* propõe dois modos de investigação dos dados. O primeiro consiste na análise segmentada dos erros e procura reduzir as fontes de erros pela tentativa de identificação da origem das incorrecções nos dados. Assim, a colecção dos valores errados e que violam regras constitui-se num potencial conjunto de análise. Através de processos de decomposição e isolamento dos dados irregulares obtêm-se os elementos dos dados que apresentam valores errados passíveis de ser observados quanto à variabilidade face aos restantes elementos da população dos dados. O segundo modo respeita à análise dos eventos dos dados que causam erros. Esta análise envolve todos os processos que captam ou transformam os dados: processos de captura dos dados, a decadência dos dados, processos de movimento e reestruturação dos dados e os pontos de conversão dos dados em informação do negócio. Este modo de investigação tende para a identificação dos processos responsáveis pela degeneração da linhagem dos dados.

Seguidamente, é enquadrado o desenvolvimento das medidas de reparação. As medidas de reparação podem ser classificadas em soluções de curto e longo prazo. As primeiras correspondem a estratégias de reparação, que se justificam por custos e tempo de implementação mais reduzidos. Por isso, estas medidas devem ser encaradas pelas organizações como as primeiras a implementar porque possibilitam a obtenção de rápidos melhoramentos (e.g. o melhoramento dos processos de aquisição dos dados, a selecção de melhores ferramentas de limpeza dos dados ou a introdução de políticas inovadoras de gestão dos recursos humanos). Enquanto, as soluções de longo prazo pretendem a resolução definitiva dos problemas com os dados (e.g. a substituição duma aplicação). O âmbito da aplicação destas soluções pode abranger parte ou a globalidade dos patamares da arquitectura num SDW.

Depois, devem ser executadas as medidas de reparação definidas. A implementação exige uma coordenação dos recursos da organização (financeiros, humanos e materiais) e por último, procede-se à monitorização dos resultados obtidos, capaz de ser conseguida por duas vias. A primeira consiste na validação dos esforços de implementação, ou seja, são avaliadas as mudanças através da medição da qualidade dos dados antes e após a ocorrência das mudanças. Por vezes, as mudanças operadas podem provocar inconvenientes não previstos, como seja a perda de desempenho na apresentação dos dados. O segundo modo de monitorização dos resultados procura aferir sobre a existência de novos problemas. A solução preconizada pode resolver apenas parcialmente os problemas anteriormente diagnosticados ou revelar outro tipo de falhas nos dados ainda não detectadas.

Em suma, a proposta preconizada por *Olson* revela uma preocupação em vista a resolução de problemas existentes e emergentes, por meio da identificação e solução dos erros nos valores dos dados. A orientação seguida traduz-se, predominantemente, numa estratégia de reparação, visando a resolução de problemas em vez de tomar uma atitude preventiva quanto a esta questão. Todavia, parece possível a coabitação entre as duas abordagens para alcançar o objectivo final de garantia duma elevada qualidade dos dados em SDWs.

### 4.2.8 A proposta de Ballou & Tayi

A proposta [Ballou & Tayi, 1999] considera que para os esforços de melhoramento dos dados num DW produzirem efeitos, os utilizadores e os gestores do DW têm de pensar de modo sistémico sobre o que é exigido. É igualmente necessário que as organizações distingam os dados vitais dos desejáveis. Esta proposta constitui-se num modelo de desenvolvimento que procura identificar projectos de qualidade dos dados que visem o aumento da utilidade dos dados num DW. O modelo assenta em negociações sistemáticas entre os diversos intervenientes, como seja o valor a obter sobre conjuntos de dados não disponíveis no momento e a avaliação dos ganhos gerados pela redução da quantidade de dados armazenados (e.g. a opção pela melhoria da frescura dos dados pode ficar a dever-se à ausência de determinados dados). Apesar da aparente aplicação do modelo, esta proposta não teve implementação no mundo real [Leitheiser, 2001]. A operacionalidade deste modelo de programação numérica exige o cumprimento de um conjunto de pressupostos da responsabilidade do gestor do DW [Ballou & Tayi, 1999] [Leitheiser, 2001]:

- Identificar e atribuir prioridades às actividades organizacionais que o DW suporta.
- Determinar os dados internos e externos necessários para suportar essas actividades.
- Avaliar a qualidade dos dados para cada conjunto de dados necessário.
- Definir os projectos potenciais para o melhoramento da qualidade dos dados no DW.
- Estimar o impacto na qualidade dos dados para cada projecto.
- Definir as alterações de utilidade do DW causadas por cada projecto.

### 4.2.9 A proposta de Helfert & Herrmann

A proposta [Helfert & Herrmann, 2002] descreve uma aproximação para a gestão da qualidade dos dados em SDWs através de metadados baseados num sistema de qualidade dos dados. A aproximação assenta na TQM e procura fornecer um método para a gestão da qualidade dos dados em SDWs através de duas tarefas chave: o planeamento da qualidade e o controlo da qualidade.

A primeira tarefa consiste na reunião das expectativas e requisitos dos consumidores de modo a convertê-los em especificações e processos de entrega dos dados. Os critérios de qualidade são seleccionados, classificados e atribuídos por graus de prioridade no processo. A segunda tarefa respeita à verificação dos processos de entrega dos dados, através da medição quantitativa da qualidade dos dados e procura assegurar que estes cumprem as especificações estabelecidas.

A investigação identifica três níveis de acção da qualidade dos dados num SDW: utilizador, produto e conceptual. O nível do utilizador respeita as exigências de qualidade dos utilizadores e por isso, representa o nível externo. Baseados nestes requisitos, torna-se possível derivar as especificações do produto e dos processos. Assim, obtém-se o nível do produto e consequentemente o nível do processo. Dada a diferente natureza dos níveis, diferentes métodos de avaliação têm de ser considerados para a medição da qualidade em cada nível. Estes três níveis de qualidade originam dois factores da qualidade: a qualidade de desenho e a qualidade de conformidade. As especificações do produto servem de ponto de partida para efectuar a avaliação da qualidade do processo de produção dos dados. Durante o processo produtivo, a qualidade de conformidade encarrega-se de observar os valores dos dados e realiza uma avaliação para averiguar da conformidade com as especificações. A figura 4-7 ilustra a ligação entre os níveis considerados e os dois factores obtidos.

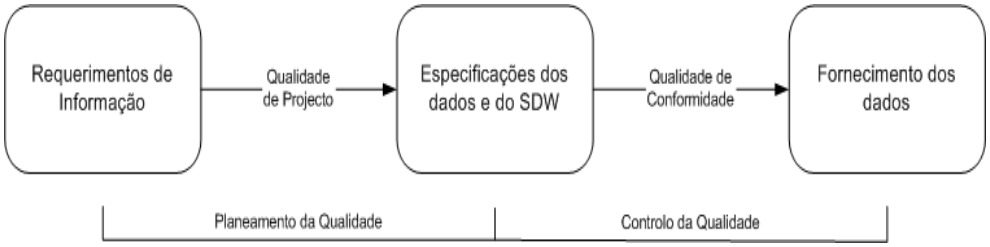



**Figura 4-7 – Factores e níveis da qualidade dos dados [Helfert & Herrmann, 2002].**

O sistema de qualidade proposto adopta como referência o melhoramento contínuo da qualidade (TQM). Para tal, a gestão dos metadados assume um papel crucial em todo o processo, em especial, os metadados relativos aos processos de transformação e esquemas dos dados.

### 4.2.10 A proposta do DWQ

O modelo de arquitectura proposto em [Jeusfeld et al., 1998] integra-se nas investigações desenvolvidas pelo DWQ e surge como forma de colmatar as dificuldades de compreensão do desenho e análise da qualidade dos DWs pelos intervenientes. O estudo fornece um método para a representação dos objectos dum DW sob uma plataforma de metadados. Os objectos (meta-objectos)

dum SDW podem resultar das diversas camadas (SO e DW) e sob as diferentes perspectivas (conceptual, lógica e física) (e.g. um objecto pode ser uma tabela do SO representada nos metadados como a definição da relação e o modelo de dados). Assim, os metadados devem descrever todos os componentes estruturantes dum DW. A ideia geral consiste em avaliar a qualidade dos objectos através da representação de um meta-modelo de qualidade. O meta-modelo de qualidade situa-se no repositório dos metadados e deve representar, claramente, os objectivos de qualidade a atingir em cada objecto, de modo a satisfazer os anseios dos diversos intervenientes. Este meta-modelo deve reunir, igualmente, as medições efectuadas sobre a qualidade. As aferições realizadas à qualidade dum DW são encaminhadas para um local anexo ao próprio DW que contém os metadados associados à qualidade. O meta-modelo fornece uma estrutura para a formulação da especificação dos objectivos a atingir, das consultas a executar e das medições a realizar ao nível da qualidade. A proposta contextualiza e adapta a aproximação GQM a ambientes de DW, gerando as seguintes diferenças [Vassiliadis et al., 1999]:

- A distinção clara entre objectivos de qualidade subjectivos dos intervenientes e os factores de qualidade objectivos associados aos objectos dum DW.
- A realização dos objectivos de qualidade resulta da avaliação dos factores de qualidade associados.
- As consultas de qualidade são implementadas e executadas no repositório de metadados.

A figura 4-8 representa o meta-modelo de qualidade. A parte superior do modelo em torno do *objectivo de qualidade* permite que os intervenientes estabeleçam os seus requisitos. Um *objecto a medir* é um objecto na perspectiva conceptual, lógica ou física do DW (e.g. um *wrapper* ou o esquema conceptual duma fonte de dados). Um *objectivo de qualidade* relaciona-se com um objecto e um interveniente e consiste nos requisitos de um interveniente sobre o *objecto a medir* (e.g. os decisores pretendem aumentar a exactidão dos valores da localidade dos alunos para 97% até ao fim do mês). Além de indicar o interveniente, o *objectivo de qualidade* é especificado sobre: o *propósito* a alcançar (e.g. aumentar, encontrar); a *descrição* do objectivo a atingir e as *dimensões da qualidade* associadas. Uma *dimensão da qualidade* é o conceito base para a formulação de *objectivos de qualidade* (e.g. oportunidade). Cada dimensão pode ser refinada pela hierarquização de sub-dimensões, orientadas por um interveniente. Deste modo, os intervenientes manifestam as suas preferências quanto às dimensões a considerar e à hierarquia entre as dimensões e as sub-dimensões associadas. Uma *consulta de qualidade* serve para medir se o objectivo de qualidade está a ser correntemente satisfeito. Ou seja, a análise dos *objectivos de qualidade* é realizada através de *consultas de qualidade* efectuadas para as *medidas de qualidade* (e.g. verificar quando

a exactidão da localidade da tabela alunos atingir os 97%). A associação entre as *medidas de qualidade* e o *objecto a medir* realça a exigência de monitorização sobre o *objecto a medir*. Uma *medição de qualidade* é o registo duma medida de qualidade sobre um *objecto a medir*. A *unidade da métrica* indica a unidade de medida de valor da qualidade (e.g. valores nulos por linha). O *domínio da qualidade* enuncia os valores possíveis de constar nos resultados das medidas. O *intervalo de qualidade* é um conjunto de valores de qualidade, que genericamente se traduz num subconjunto do *domínio da qualidade* [Vassiliadis et al., 1999] [Jeusfeld et al., 1998].

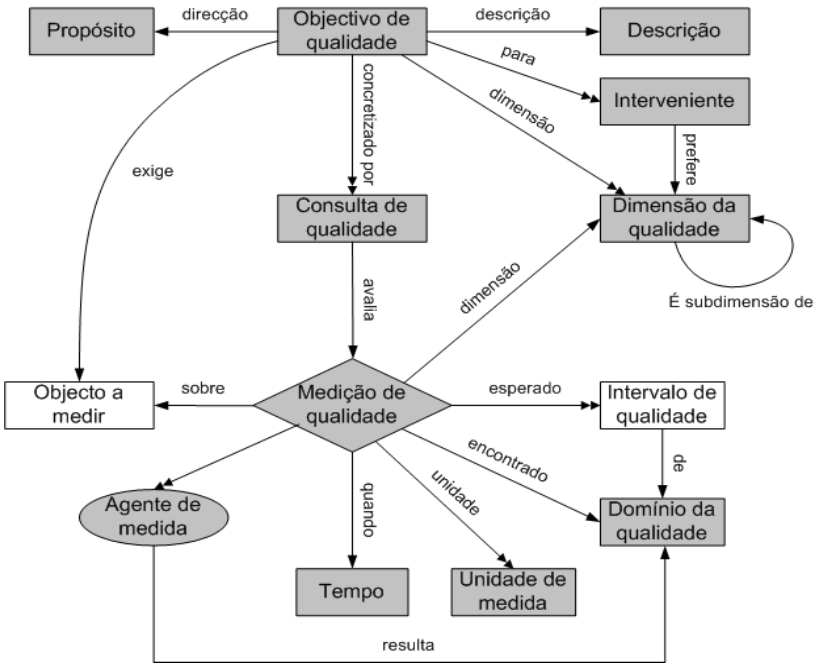



**Figura 4-8 – O meta-modelo da qualidade [Jeusfeld et al., 1998].**

## 4.2.11 A proposta de Vassiliadis

A investigação desenvolvida em [Vassiliadis, 2000] segue o trajecto trilhado pelo DWQ e assenta num modelo de qualidade claramente influenciado pela proposta anteriormente apresentada. A investigação considera a gestão do ciclo de vida dum DW assente na vigilância de três ópticas fundamentais objectivamente circunscritas e agindo suplementarmente entre elas: a arquitectura, os processos e a qualidade. A óptica da arquitectura centra-se na compreensão dos componentes estáticos dum DW (e.g. as fontes de dados). A óptica dos processos visa apreender os componentes dinâmicos inerentes a um DW. Por último, a óptica da qualidade pretende aferir o nível de qualidade proporcionada por um DW tendo em vista a satisfação dos desejos dos intervenientes. As

vertentes consideradas devem ser sustentadas num modelo capaz de descrever fiel e coerentemente as três áreas.

Contextualizando ao tema da dissertação, iremos debruçar-nos sobre a óptica da qualidade. O meta-modelo de qualidade proposto está descrito nos metadados dum DW e encontra-se disposto sob três camadas de instanciação: a camada do meta-modelo, a camada dos metadados e a camada dos valores. A camada do meta-modelo segue o modelo GQM, mas adapta-o a ambientes de DW. Quanto à camada de metadados, são definidos objectivos de qualidade específicos para situações particulares num SDW. Por fim, a camada dos valores reais define os caminhos de medida no mundo real. A figura 4-9 ilustra o meta-modelo da qualidade.

Um *objectivo de qualidade* é um assunto de qualidade sobre um DW que um interveniente pretende operar (e.g. a eliminação de valores nulos na coluna nome da tabela de alunos). As *dimensões da qualidade* usam-se para definir abstractamente aspectos particulares da qualidade dos dados que são percepcionados pelos intervenientes (e.g. completude). As dimensões podem ser especificadas pelos factores de qualidade directamente influentes na sua obtenção e orientadas segundo os diferentes intervenientes. Um *objectivo de qualidade* é definido operacionalmente por um conjunto de *questões de qualidade.* Cada questão está ligada a *métricas de qualidade* ou factores de qualidade concretos para efectuar medições de qualidade. Uma *métrica* é definida para um objecto do DW e compreende valores expectáveis e aceitáveis, valores medidos, registo de tempo, etc..

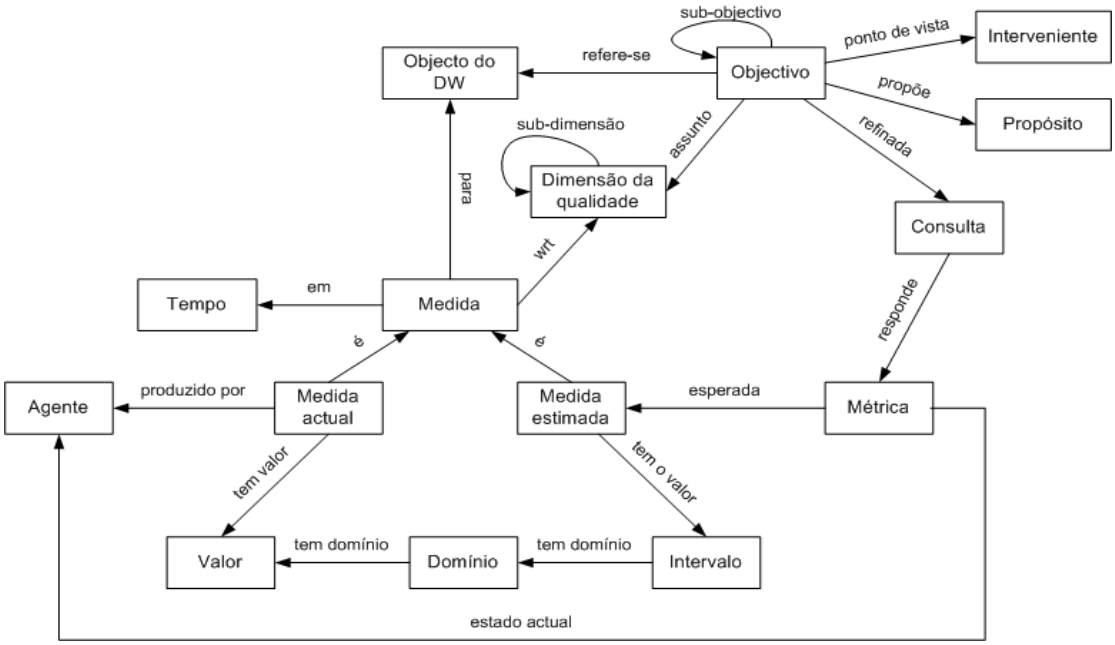



**Figura 4-9 – O meta-modelo da qualidade [Vassiliadis, 2000].**

### 4.2.12 Comparação das propostas

As abordagens apresentadas visam a melhoria da qualidade dos dados em SDWs. Porém, as propostas apresentam algumas diferenças que permitem complementá-las entre si. A tabela 4-2 procura estabelecer a comparação entre as propostas apresentadas, em especial, no que concerne à especificidade a ambientes de DW e à política de correcção dos defeitos.

| Proposta | Tema | Específico para SDW | Descrição | Política | Origem |
|---|---|---|---|---|---|
| [Redman, 1995] | Estratégia para a melhoria da qualidade dos dados | Não | Aproximação que descreve três estratégias de melhoramento da qualidade dos dados: identificação do problema, o tratamento dos dados como assuntos de negócio e a implementação de sistemas de qualidade mais avançados. | Reparação | |
| QFD [Helfert & Radon, 2000] [Vassiliadis, 2000] | Método de planeamento da qualidade da produção | Não | Método de planeamento da qualidade da produção, que se baseia nos requisitos dos consumidores e desenvolve pontos-chave referentes à garantia da qualidade ao longo do processo produtivo. | Prevenção | |
| TDQM [Wang, 1998] | Metodologia para a qualidade dos dados | Não | Metodologia para a avaliação da qualidade dos dados em sistemas de informação. Apresenta o conceito de produto informação e estabelece o ciclo: definir, medir, analisar e melhorar | Prevenção | TQM<br>Dimensões de [Wang et al., 1994] |
| IPMAP [Shankaranarayan et al., 2000] | Representação do processo produtivo | Não | Modela o processo produtivo e compreende o modo como os vários componentes do produto informação agem em conjunto. Paralelamente, a etapa de avaliação visa prevenir, detectar ou corrigir algumas irregularidades dos dados. Possibilita estimar o tempo de entrega, o rastreio e o alcance dos processos. | Prevenção | TDQM [Wang, 1998]<br>Information Systems Manufacturing [Ballou et al., 1998] |
| [Shankaranarayan, 2005] | Implementação da TDQM e IPMAP em SDWs | Sim | Aproximação para a gestão da qualidade dos dados em SDWs através da implementação da TDQM. A proposta recorre ao modelo IPMAP para a representação gráfica do sistema de produção dos dados. Integra também os metadados sobre a qualidade dos dados aos metadados existentes num DW. | Prevenção | TDQM [Wang, 1998]<br>IPMAP [Shankaranarayan et al., 2000] |
| [Olson, 2003] | Estratégia para a melhoria da qualidade dos dados | Não | Estratégia de melhoria da qualidade dos dados assente numa sequência de passos: conversão de factos em assuntos, avaliação do impacto, investigação das causas, desenvolvimento de medidas de reparação, execução das medidas de reparação aprovadas e a monitorização dos resultados obtidos. | Reparação | |
| [Ballou & Tayi, 1999] | Modelo de programação numérica | Sim | Proposta de um modelo de programação numérica para a identificação de projectos de qualidade dos dados que visem o aumento da utilidade dos dados num DW. O modelo incorpora um sistema de negociações sistemáticas entre os intervenientes. | Prevenção | |

**Tabela 4-2 – Tabela resumo de metodologias e modelos adoptados para a melhoria da qualidade dos dados em SDWs (continua).**

| Proposta | Tema | Específico para SDW | Descrição | Política | Origem |
|---|---|---|---|---|---|
| [Helfert & Herrmann, 2002] | Gestão da qualidade dos dados | Sim | A proposta visa a melhoria da qualidade dos dados em SDWs através do recurso a metadados de sistema de qualidade dos dados. A proposta recorre aos princípios da TQM e a dois conceitos chave: qualidade de conformidade e qualidade de projecto. | Prevenção | TQM |
| TIQM [English, 2003] | Metodologia para a qualidade dos dados | Não | Metodologia que institucionaliza uma atitude organizacional para a melhoria da qualidade dos dados. Assenta nos catorze pontos de *Deming* e é composta por seis processos que promovem a melhoria contínua da qualidade dos dados. | Prevenção | TQM |
| DWQ [Jeusfeld et al., 1998] | Meta-modelo da qualidade | Sim | Define um meta-modelo para avaliação da qualidade dos objectos do SDW, tendo em conta os objectivos propostos e orientados pelos requisitos dos intervenientes. | Prevenção | |
| [Vassiliadis, 2000] | Meta-modelo da qualidade | Sim | Define um meta-modelo para avaliação da qualidade dos objectos do SDW, tendo em conta os objectivos propostos e orientados pelos requisitos dos intervenientes. | Prevenção | |

**Tabela 4-2 – Tabela resumo de metodologias e modelos adoptados para a melhoria da qualidade dos dados em SDWs (continuação).**

## 4.3 Plataforma do sistema de qualidade dos dados em SDWs

A plataforma proposta, de sistema de qualidade dos dados, é determinada pelo fluxo de circulação dos dados num SDW. A proposta procura abarcar as tarefas comummente designadas de *Back End* [Chaudhuri & Dayal, 1997]. Estas tarefas justificam a "fatia de leão" do esforço dispendido na implementação e manutenção de um DW (meios humanos, técnicos, tecnológicos, financeiros e temporais). Normalmente, a natureza destes processos tende para a automatização, com a participação pontual de peritos nas questões relacionadas com os dados. Logo, é importante tomar consciência que a complexidade e quantidade de anomalias nos dados dificulta a automatização do processo de limpeza dos dados [Galhardas et al., 2001].

A estratégia seguida opta por uma atitude preventiva da ocorrência de irregularidades dos dados num SDW. Assim, pressupõe-se atingir num primeiro momento, a correcção dos erros dos dados e num momento posterior, a adopção duma postura que preveja a melhoraria dos processos para prevenir a recorrência de erros [English, 2004]. Ainda à luz de *English*, prevê-se que a estratégia de melhoramento dos dados se inicie pelos resultados fornecidos pelas análises dos dados, pela aplicação de métricas que indiquem os problemas mais prementes ou ainda pela adopção do conceito de atributos críticos, ou seja, aqueles atributos que, pela sua qualidade débil, originam graves repercussões no fluxo circulatório dos dados num SDW [Inmon et al., 1998].

A plataforma considera cinco pontos de circulação dos dados num SDW, conforme podemos verificar na figura 4-10: os dados existentes nas fontes que são relevantes para os decisores (FD1, FD2 e FD3); os dados extraídos das fontes; os dados constantes na ARD; os dados a carregar no DW e os dados constantes no DW. Cada zona de circulação dos dados configura um conjunto de questões problemáticas particulares sobre os dados e carentes de tratamento específico.

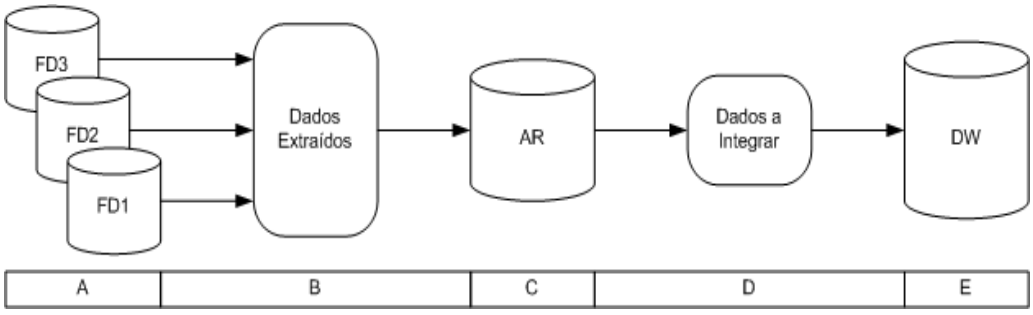



**Figura 4-10 – Plataforma do sistema de qualidade dos dados em SDWs.**

A plataforma do sistema de qualidade prevê uma atitude de gestão dos dados encarada como um processo contínuo que requere persistência e vigilância. Nesse sentido, a gestão dos dados deve, igualmente, assentar num conjunto de métricas distribuídas estrategicamente pelo fluxo circulatório dos dados nos SDWs. As métricas devem ser associadas e geridas através de actividades de monitorização, de recolha, de sintetização e de distribuição desses instrumentos de medida. O objectivo dessas medições consiste em possuir uma aferição sobre a qualidade dos dados de forma factual em detrimento de uma avaliação percepcional. A recolha dos dados fornecidos pelas métricas deve ser confrontado com os requisitos definidos pelos utilizadores dos dados de modo a decidir sobre as iniciativas correctivas a implementar [Gonzales, 2004].

As técnicas e ferramentas devem ser aplicadas de modo integrado e respeitando um plano global em vista o melhoramento da qualidade dos dados nos SDWs. Algumas investigações congregam todas as propostas como etapas sequenciais pertencentes ao processo de limpeza dos dados [Müller & Freytag, 2002] [Rahm & Do, 2000]. Todavia, a abordagem usada na presente dissertação considera as propostas dispostas em categorias de aplicação, de forma a revelar a sua área de actuação concreta (de acordo com a arquitectura dum SDW), a crescente especialização dos diversos domínios e em consequência a importância vital da gestão global da qualidade dos dados em SDWs. Interessa ter presente que o grau de qualidade dos dados a atingir, ou seja, as deficiências a corrigir, deve corresponder às especificações previamente estabelecidas e que resultam da negociação entre as diversas partes interessadas [Müller & Freytag, 2002]. Este factor é relevante porque o desempenho da execução destas tarefas (qualidade de conformidade) é directamente influenciado pelas especificações acordadas (qualidade de projecto).

A compreensão destas propostas deve pressupor um entendimento alargado sobre o modo como os dados devem ser geridos em SDWs, bem como a especialização dos processos de gestão dos dados. Assim, pretendemos enunciar uma sequência dos passos necessários num ciclo capaz de, forma eficaz e eficiente, realizar uma gestão dos dados em ambientes de DW: a análise, a transformação e limpeza, a integração, o enriquecimento e a monitorização dos dados.

### 4.3.1 Zona A: fontes de dados

As propostas a integrar nesta zona compreendem, essencialmente, as ferramentas de análise e auditoria aos dados e visam detectar os tipos de defeitos existentes nos dados para posterior resolução. A constatação sobre os defeitos a solucionar permite aferir sobre o verdadeiro estado dos dados constantes no SO. A determinação do nível de qualidade dos dados pressupõe conhecer as exigências de informação da organização. Uma consistente qualidade dos dados é o estado do

repositório de dados em que a qualidade dos dados é sobejamente compreendida e a qualidade do repositório dos dados conhecida [Brackett, 1996].

Um importante resultado oriundo da análise e avaliação dos dados consiste na identificação de todos os atributos críticos e assim, obter uma lista dos atributos prioritários. A associação de prioridades aos atributos serve como linha orientadora no modo de resolução das anomalias, indicando o local onde a introdução de padrões de qualidade deve ser realizada. Por exemplo, um valor pode ser considerado indispensável na concretização dum PI, o seu grau de susceptibilidade ao erro ser elevado e o tempo de captura ser único, o que contribui para uma solução na fonte que colmate a ausência de deficiências na recolha e armazenamento dos valores.

As soluções de melhoramento dos dados, integradas nesta categoria, visam a obtenção de informações ou indicadores sobre as características e problemas dos dados e devem ser corporizadas através de metadados. Em seguida, são apresentadas duas aproximações que podem ser classificadas nesta categoria: *data profiling* e auditoria dos dados.

## Data profiling

A aplicação de *data profiling* no SO justifica-se pela obtenção de informações relativas aos dados existentes nas bases de dados. Em [Olson, 2003] define-se *data profiling* como a aplicação de técnicas de análise com o propósito de conhecer o conteúdo, a estrutura, e a qualidade dos dados actuais. A existência de metadados que denotem escassez de elementos sobre os esquemas ou os dados armazenados nos sistemas de dados do SO inviabilizam um conhecimento sobre os problemas com esses dados. É importante analisar os valores existentes tendo em vista a obtenção de metadados sobre as características ou os padrões existentes nos dados. Estes metadados são essenciais na detecção dos defeitos. Além disso, é igualmente útil a disponibilização de metadados e informações de qualidade em projectos DW. Estes projectos implicam a transferência dos dados, provenientes do SO, para outras estruturas de dados [Rahm & Do, 2000]. Portanto, a técnica de *data profiling* procura obter informações reais sobre os dados e para isso, baseia-se na assumpção que os metadados nas fontes ou são incompletos ou estão errados.

Em [Olson, 2003] é previsto um modelo geral assente numa componente de fornecimento de informações ao processo (figura 4-11): os metadados disponíveis e os dados. Os metadados existentes, mesmo quando incompletos, são importantes porque fornecem informações relativas à estrutura básica mínima dos dados. Os resultados gerados pelo processo de *data profiling* direccionam-se em duas vias. Por um lado, a obtenção de metadados mais correctos, mais descritivos, mais ricos e servindo de precioso auxilio nas actividades de ETL dum SDW. Por outro lado, em

factos que evidenciem as discrepâncias entre os dados e os metadados correctos. Estes factos mostram as deformidades dos dados e tornam-se num ponto de partida para a investigação das causas dos erros verificadas.

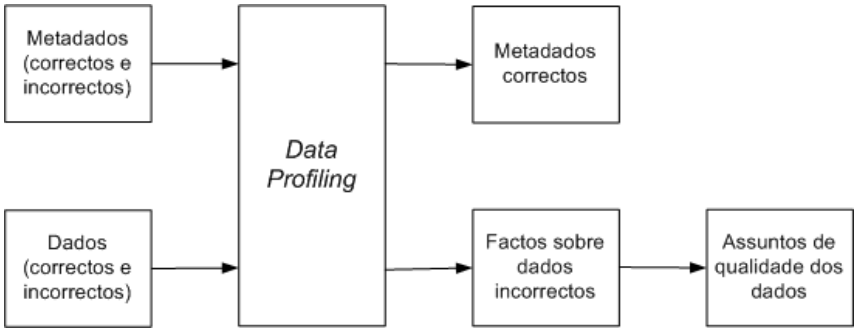



**Figura 4-11 – Modelo de Data *Profiling* [Olson, 2003].**

O conhecimento sobre a realidade dos dados existentes no SO é condição necessária para uma posterior limpeza dos dados a integrar num DW. A análise dos dados das fontes permite aferir sobre as diversas irregularidades verificadas e os níveis de gravidade dessas irregularidades. Em [Kimball, 2004] é referido que os valores obtidos pelo *data profiling* determinam o rumo a seguir na implementação dum DW. A primeira hipótese consiste em abandonar o projecto de DW porque as fontes impedem a disponibilização dos dados exigidos pelos decisores. A segunda hipótese condiciona o sucesso dum DW, a alterações prementes nas fontes. Por fim, permite determinar as melhores soluções a implementar no momento das operações de limpeza dos dados na ARD.

A existência de metadados reveladores dos perfis das fontes de dados mostra-se, igualmente, interessante sobre as perspectivas de actualidade dos dados num DW. Assim, além das preocupações relativas às incorrecções dos dados (dimensão correcção), outras necessidades, como a frescura dos dados podem ser pertinentes. A determinação do perfil de frescura das fontes de dados é determinante para as exigências de actualidade dos dados no DW. É, por isso, importante a manutenção de metadados que providenciem a frequência de actualização das fontes (e.g. o tempo da última actualização ou a frequência de actualizações). A contingência apresentada releva a urgência em possuir perfis das fontes de dados adequados a uma manutenção dos dados actualizados no DW [Bouzeghoub & Peralta, 2004].

## Auditoria dos dados

Ainda no plano de soluções para a análise dos dados existentes no SO é possível recorrer a técnicas de auditoria dos dados como forma de debelar imperfeições dos dados e proporcionar a obtenção de dados mais puros. As aplicações de auditoria caracterizam-se por incorporarem algorit-

mos de mineração para medir e melhorar a qualidade dos dados [Jarke et al., 2003]. Alguns dados apresentam impurezas, nomeadamente em termos de ruído (valores errados ou valores aberrantes) e de ausência de valores. As atitudes correctivas que tentam colmatar estas deficiências podem passar, sinteticamente, pelo preenchimento de dados ausentes, pelo alisamento do ruído e pela identificação de valores aberrantes e inconsistentes [Carvalho, 2003].

No que respeita ao preenchimento de valores ausentes, os tratamentos usuais podem passar pela inserção do valor da média ou moda simples ou por classe, ou por usar o valor mais provável segundo um modelo (regressão, regra de Bayes, árvores de decisão). Relativamente ao ruído apresentado pelos dados, algumas técnicas, como o alisamento e a regressão permitem a remoção de ruído. O alisamento consiste em distribuir os dados ordenados, tendo como referência os seus vizinhos. Quanto à identificação de valores aberrantes é possível recorrer a técnicas de segmentação (figura 4-12) ou de regressão linear (figura 4-13).

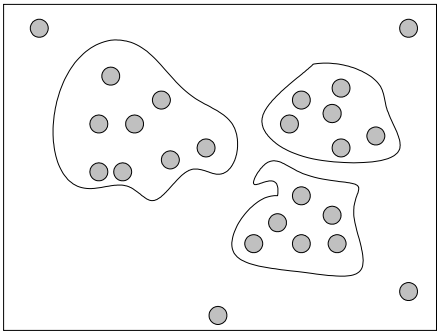


**Figura 4-12 – Técnicas de segmentação para eliminação de valores aberrantes.**

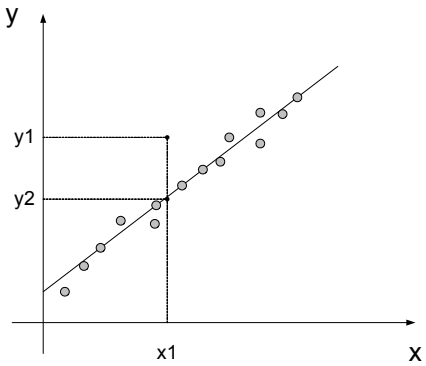



**Figura 4-13 – O alisamento dos dados com recurso à regressão linear.**

As técnicas de mineração permitem, igualmente, descobrir padrões dos dados em conjuntos alargados de dados. Esta capacidade pode revelar-se bastante útil na identificação de relacionamentos entre diversas colunas duma mesma entidade [Rahm & Do, 2000].

### 4.3.2 Zona B: dados extraídos das fontes

Os dados extraídos do SO encontram-se em trânsito para a ARD e são as matérias-primas essenciais para a elaboração dos PIs. Estes dados em bruto são submetidos a um conjunto de operações, comummente designadas, de transformação e limpeza dos dados. A limpeza dos dados visa detectar e remover anomalias com o objectivo de aumentar e melhorar a qualidade destes [Rahm & Do, 2000]. Em [Müller & Freytag, 2002] define-se a limpeza de dados como um conjunto de operações executadas, em vista a remoção de imperfeições e consequentemente, a obtenção de um repositório de dados mais fielmente representativo do mundo real. A limpeza comum dos dados segue uma abordagem de inspecção porque o principal objectivo consiste na identificação e remoção dos problemas nos dados após estes terem sido criados [Amaral et al., 2002].

Estas operações são executadas individualmente em cada fonte de dados e correspondem às iniciativas clássicas de limpeza dos dados (e.g. a correcção de erros sintácticos e a preparação dos dados para integração futura). Os métodos geralmente utilizados nesta fase são: a decomposição dos dados para obter elementos atómicos; a estandardização, correcção e normalização dos dados; o preenchimento de valores ausentes, a aplicação das regras de integridade referencial e o enriquecimento do conteúdo dos dados. A resolução do problema dos valores duplicados nos dados das fontes é, também, um outro método de aplicação. Porém, abordaremos este assunto na zona de reunião dos dados provenientes das fontes porque se trata, provavelmente, da operação de limpeza mais delicada executada nessa zona [Rahm & Do, 2000].

#### Decomposição dos dados

Esta operação consiste na separação dos valores dos dados em componentes atómicos [English, 2004]. Para [Müller & Freytag, 2002], a decomposição dos dados é executada para a detecção de erros de sintaxe. Alguns campos de dados caracterizam-se por capturarem múltiplos valores individuais que, uma vez dissolvidos, permitem encontrar uma representação mais precisa e usável nas tarefas de limpeza posteriores (validação, estandardização, correcção e eliminação de valores duplicados) [Rahm & Do, 2000]. Em [Olson, 2003] refere-se a este tipo de problemas, como o sobrecarregamento dos dados (e.g. a inclusão do caracter '#' no nome dum aluno indica que faleceu). Geralmente, os dados para os quais existem diversas ferramentas de tratamento específicas respeitam aos campos nome e morada (tabela 4-3).

| Dados Originais | Decomposição por Elementos | |
|---|---|---|
| 1999 Herdade da Erva Vinho Tinto Alentejano VQPRD | Marca | Herdade da Erva |
| | Ano | 1999 |
| | Tipo | Tinto |
| | Região | Alentejo |
| | Categoria | VQPRD |

**Tabela 4-3 – Decomposição dos dados relativos à designação de um produto.**

A decomposição dos campos em elementos atómicos possibilita, num primeiro momento, a validação e correcção de valores dos dados, através da comparação de valores similares (e.g. a separação de uma morada em nome de rua e número permite a detecção de outras cadeias de caracteres semelhantes). Num momento posterior, é possível enriquecer as linhas de dados com informações complementares, baseadas nos valores decompostos individualmente ou em conjugação. A ausência de alguns valores pode por este meio ser colmatada (e.g. a obtenção da totalidade de um código postal após o fraccionamento da morada). Por último, a decomposição dos dados é uma parte importante do processo de verificação porque permite validar outros campos no registo (e.g. a separação do nome de um aluno possibilita confrontar o sexo desse aluno) [English, 2004].

## Estandardização e normalização dos dados

A estandardização dos dados resume-se a conversões operadas nos dados para um formato uniforme definido para um DW. A uniformização dos dados deve ser um facto presente nos mais variados tipos de dados porque facilita a sua integração e a resolução de conflitos (e.g. os dados de texto devem ser condensados e uniformizados pela remoção de sufixos e prefixos, da remoção de sinónimos e do estabelecimento de abreviaturas de modo consistente) (tabela 4-4).

| Dados Originais | Standard Escolhido | Dados Estandardizados |
|---|---|---|
| Coop. Agrícola do Alentejo | | Cooperativa Agrícola do Alentejo |
| CAA | | Cooperativa Agrícola do Alentejo |
| C.A.A. | | Cooperativa Agrícola do Alentejo |
| Coperativa A. Alentejo | | Cooperativa Agrícola do Alentejo |
| Cooperativa Agrícola do Alentejo | Cooperativa Agrícola do Alentejo | Cooperativa Agrícola do Alentejo |
| C. A. Alentejo | | Cooperativa Agrícola do Alentejo |
| Cooperativa Agrícola Alentejana | | Cooperativa Agrícola do Alentejo |

**Tabela 4-4 – Estandardização dos dados relativos a um nome.**

Este tipo de operações revela-se particularmente útil aquando das interrogações sobre os dados (cf. tabela 4-4). A discrepância entre os resultados obtidos por uma mesma consulta: *SELECT * FROM clientes WHERE nome_cliente = 'Cooperativa Agrícola do Alentejo'*, antes e após a estandardização dos dados torna evidente a necessidade destas transformações nos dados.

As técnicas de normalização dos dados possibilitam a definição de padrões regulares no formato dos dados constantes nas bases de dados (e.g. número de contribuinte e telefone). Os números de contribuinte podem ser normalizados de acordo com um formato previamente estabelecido, como seja 999 999 999. A aplicação deste padrão aos valores registados permite converter os diferentes formatos dos números existentes num único formato [English, 2004].

### Validação e correcção dos dados

Esta etapa pretende examinar e corrigir os valores errados capturados em cada fonte de dados. É comum o recurso a dicionários de sinónimos e à verificação ortográfica para a identificação e reparação de irregularidades. Alguns dados apresentam-se aparentemente correctos e estandardizados, mas na realidade denotam conflitos ou contradições entre colunas duma mesma linha (e.g. o código postal não corresponde à morada associada ou a data de nascimento de um aluno não respeita à idade registada). Estas operações compreendem, igualmente, a rectificação de valores que violam o domínio de valores definido para a coluna, as regras do negócio e o preenchimento de valores ausentes ou incompletos [English, 2004] [Rahm & Do, 2000]. Algumas formas de preenchimento de valores nos dados foram já referidas anteriormente, como seja, pela aplicação de técnicas de auditoria ou de mineração dos dados.

### Regras de integridade

A aplicação das regras de integridade descreve o problema em assegurar a integridade dos dados após a ocorrência de operações nos repositórios dos dados (e.g. inserção, eliminação ou actualização de registos). Em [Müller & Freytag, 2002] são feitas referências a duas propostas de tratamento desta problemática: a verificação das regras de integridade e a manutenção das regras de integridade. A primeira aproximação rejeita as transacções que podem violar alguma regra de integridade dos dados. A segunda aproximação ocupa-se da adição ou actualização de dados para que o repositório de dados não viole as regras de integridade.

## 4.3.3 Zona C: dados estacionados na ARD

Após a limpeza dos dados em cada uma das fontes, as operações devem centrar-se na integração dos esquemas e dados e na remoção de problemas pela reunião de várias fontes. As tarefas apli-

cadas nesta zona são em tudo semelhantes às executadas em cada uma das fontes. Todavia, neste local há particular incidência de umas operações (e.g. remoção de duplicados) em detrimento de outras (e.g. decomposição dos dados). A antecipação de algumas tarefas facilita e alivia a complexidade e o número de operações a realizar na ARD.

Recorremos a um exemplo que evidencia a necessidade de procedimentos de limpeza e transformação dos dados e esquemas das tabelas na ARD (tabelas 4-5 e 4-6). Conforme é dado a observar, verifica-se que não foram tomadas nenhumas iniciativas prévias de tratamento dos dados nas fontes. Apesar desta ser uma situação perfeitamente plausível de ocorrer (e.g. a inclusão duma fonte de dados exterior à organização), pensamos que não corresponde à melhor abordagem na limpeza dos dados num SDW porque pressiona a ARD, um local que já concentra um alargado número de operações complexas normalmente executadas durante janelas de oportunidade reduzidas. O exemplo considerado mostra claramente alguns conflitos dos esquemas e dos dados associados. Ao nível do esquema verifica-se o conflito entre os nomes das colunas (sinónimos) e conflitos na estrutura de algumas colunas (diferentes representações do nome e morada). Enquanto que, ao nível das linhas existem diferenças na representação do sexo e na ocorrência de linhas duplicadas.

**Cliente (fonte de dados 1)**

| NC | Nome | Rua | Cidade | Sexo |
|---|---|---|---|---|
| 11 | Cristina Almeida | R. de Sant'ana à Lapa, nº 210 | 4330 Lisboa | 0 |
| 24 | José F. Lopes | Av. Vinte e Quatro, num. 65 r/c esq. | Faro | 1 |

**Tabela 4-5 – Fonte de dados 1.**

**Fregueses (fonte de dados 2)**

| ID | Último Nome | Primeiros Nomes | Morada | Sexo |
|---|---|---|---|---|
| 24 | Lopes | João F. | Av. da Infantaria, n. 6 Porto | Mas |
| 49 | Almeida | Cristina J. | Rua Santana à Lapa, 210 1350-290 | Fem |

**Tabela 4-6 – Fonte de dados 2.**

A resolução destes problemas passa, em primeiro lugar, pela integração dos esquemas e depois pela limpeza dos dados. Porém, reforçamos a necessidade de algumas das técnicas de limpeza (apresentadas anteriormente) serem executadas antes da recolha dos dados para a ARD. Uma possível solução encontra-se ilustrada na tabela 4-7.

**Clientes**

| CC | Apelido | Nomes | Morada | Cidade | C. Postal | Sexo | NC | ID |
|---|---|---|---|---|---|---|---|---|
| 1 | Almeida | Cristina J. | R. de Sant'ana à Lapa, 210 | Lisboa | 1350-290 | F | 11 | 49 |
| 2 | Lopes | João F. | Av. da Infantaria, 6 | Porto | | M | | 24 |
| 3 | Lopes | José F. | Av. 24, 65 r/c esq. | Faro | | M | 24 | |

**Tabela 4-7 – Junção das fontes de dados 1 e 2.**

A antecipação do tratamento dos dados deve ser concretizada tanto mais cedo, quanto as fontes de dados sejam internas à organização, a janela de oportunidade for manifestamente reduzida, os esquemas das fontes denotem conflitos de vária ordem e os dados sejam muito susceptíveis de denotarem imperfeições. As principais tarefas a realizar neste local são assim, as transformações ao nível do esquema e dos dados e na remoção de linhas duplicadas.

## Definição das transformações nos dados

O processo de transformação dos dados consiste, geralmente, num conjunto de etapas que incidem nas transformações ao nível do esquema ou das ocorrências dos dados. O número de fases ou operações a realizar varia de acordo com a quantidade de fontes de dados a integrar e com os níveis de anomalias dos dados a corrigir. As operações de transformação podem incidir no esquema ou nos dados a integrar, implicando a migração dos dados para um esquema destinatário. Este esquema compreende o mapeamento entre o esquema do SO e o DW, conforme ilustra a tabela 4-8, a uniformização dos tipos de dados.

| Sistema | Sexo | Dados Originais | Dados Transformados |
|---|---|---|---|
| 1 | Masculino | Mas | M |
| 1 | Feminino | Fem | F |
| 2 | Masculino | 1 | M |
| 2 | Feminino | 0 | F |

**Tabela 4-8 – Uniformização de tipos de dados.**

As causas das anomalias nos dados devem ser bem compreendidas, através da acção de ferramentas de análise dos dados, de modo a iniciar as medidas correctivas mais adequadas. Assim, implicitamente, o processo de transformação dos dados impõe grandes quantidades de metadados, como sejam, ao nível dos esquemas, das características dos dados, dos mapeamentos de transformação, da frescura e completude das fontes de dados.

## Eliminação de duplicados

A tarefa de eliminação de duplicados ocorre geralmente após a maioria das outras operações de transformação e de limpeza terem ocorrido, em especial, depois de reparadas as incorrecções e os conflitos de representação em cada fonte de dados. A eliminação de linhas duplicadas pressupõe a identificação de linhas semelhantes que representem a mesma entidade no mundo real e posteriormente, a fusão dessas linhas numa única capaz de agregar, sem redundância, as colunas relevantes e representativas da entidade considerada [Rahm & Do, 2000].

Esta questão é um dos maiores problemas verificados ao nível da qualidade dos dados e que resulta após a reunião de dados provenientes de diversas fontes. Pode ser observada, num primeiro instante, sobre as linhas constantes em cada fonte de dados e num instante posterior, na integração dos dados provenientes das diversas fontes e localizados na ARD. Neste último caso, o encontro de dados provenientes de fontes heterogéneas e geograficamente dispersas justifica a necessária conferência entre linhas iguais ou semelhantes existentes em cada uma das fontes. Geralmente, as técnicas utilizadas recorrem à confrontação de registos baseada em critérios de semelhança, pela comparação de valores em determinadas colunas (e.g. o nome, a morada ou o contacto) (tabelas 4-9 e 4-10).

| **Antes** | | **Depois** | |
|---|---|---|---|
| **Organização** | **Contacto** | **Organização** | **Contacto** |
| Metalurgia Industrial | João C. Silva | Metalurgia Industrial | João C. Silva |
| Metalurgia Ind. | J. C. Silva | Metalurgia Industrial | João C. Silva |
| Primeiro Banco de Lx. | António P. M. Pedro | Primeiro Banco de Lisboa | António P. Pedro |
| 1º Banco de Lisboa | António Pedro | Primeiro Banco de Lisboa | António P. Pedro |

**Tabela 4-9 – Confronto de registos por nome da empresa e contacto.**

| **Antes** | | **Depois** | |
|---|---|---|---|
| **Organização** | **Morada** | **Organização** | **Morada** |
| Cerâmicas Luz & Dias, Lda. | Rua da Olivença, 123 | Cerâmicas Luz & Dias, Lda. | R. da Olivença, 123 |
| Cerâmicas Luz e Dias | R. da Olivença, nº 123 | Cerâmicas Luz & Dias, Lda. | R. da Olivença, 123 |
| Paulo Sá Consultores | Av. da Armada, 2 4ºesq. | Paulo Sá Consultores | Av. da Armada, 2 4ºesq. |
| Paulo M. Sá Consultor | Avenida Armada, nº2 4ºE | Paulo Sá Consultores | Av. da Armada, 2 4ºesq. |

**Tabela 4-10 – Confronto de registos por nome da empresa e morada.**

Este tema tem sido alvo de estudo em diversas investigações, sendo o problema da eliminação de valores duplicados nas bases de dados [Low et al., 2001], também referenciado como o problema da identificação do objecto [Galhardas et al., 2000] ou o problema da fusão e remoção [Hernandez

& Stolfo, 1998], ou ainda, o problema da eliminação de linhas ligeiramente duplicadas [Ananthakrishna et al., 2002].

### 4.3.4 Zona D: dados em trânsito para o DW

Neste local, os dados encontram-se em trânsito em vista o seu carregamento e integração no DW. As iniciativas ao nível da qualidade dos dados consistem, basicamente, em pequenas transformações em cada uma das tabelas e na auditoria dos dados a integrar no DW. Estas transformações são, na sua maioria, preparações dos dados para a sua utilização adequada no contexto organizacional. Geralmente, procede-se à derivação de novas colunas baseadas nas existentes e à condensação e desagregação de valores que facilitem a visualização de tendências e estatísticas. Por seu turno, as tarefas de auditoria tentam aferir, acerca do cumprimento das regras definidas pela organização, sobre os dados a integrar.

### 4.3.5 Zona E: dados residentes no DW

O repositório do DW apresenta-se como o último bastião impeditivo da passagem de dados irregulares para os utilizadores finais [18]. Os sucessivos carregamentos de dados num DW podem inviabilizar ou desvirtuar os dados outrora consistentes e integrados. A qualidade dos dados em SDWs degrada-se com a operação do sistema e pelos incrementos de dados no DW. O recurso a técnicas de análise dos dados, como as aplicadas ao SO (*profiling* e auditoria dos dados), permite a obtenção de conhecimentos relativos aos dados constantes no DW. A utilização destas técnicas associadas a métricas sobre a qualidade dos dados proporciona a monitorização contínua sobre os valores dos campos e permite um reconhecimento imediato sempre que a qualidade dos dados descer para valores inferiores aos considerados aceitáveis (standards predefinidos).

A manutenção de metadados que possibilitem a consulta de informações relativas aos dados e aos carregamentos destes no DW permite, igualmente, aferir sobre eventuais tendências dos dados. A observação de tendências dos dados procura comparar as características de um dado carregamento com as características de carregamentos anteriores. Deste modo, é possível a rejeição de um lote de dados carregados no DW que denote discrepâncias, em termos de qualidade dos dados, relativamente aos dados já residentes no repositório.

## 4.4 Ferramentas de gestão dos dados

Após a definição e localização das anomalias detectadas nos dados que circulam nos SDWs e dos métodos usados para a resolução desses problemas, interessa enunciar algumas ferramentas capazes de auxiliarem as actividades para uma gestão coerente e preventiva dos dados. As ferramentas de remoção de deformidades nos dados devem promover, em primeiro lugar, a automatização dos processos em vista a limitação da intervenção humana e em segundo lugar, a resolução das irregularidades em cada fonte de dados e posteriormente aquando da integração de múltiplas fontes. Neste sentido, as ferramentas enunciadas tentam cobrir as actividades consideradas cruciais na gestão dos dados em SDWs: análise ou auditoria; transformação e limpeza; integração, preparação e monitorização dos dados.

As preocupações com a qualidade dos dados e a sua gestão estão agora a iniciar-se [Wang, 2004] e como tal, as investigações e ferramentas existentes representam ainda o início desta nova, o que explica a sua escassez e eficiência [Lee et al., 2000a]. Todavia, tem-se assistido a inúmeros progressos, tanto a nível académico como a nível dos principais fabricantes de *software* nesta área [8] [10] [14] [15]. Diversas investigações e estudos tratam a classificação das propostas de tratamento dos dados de diferentes modos: a distinção entre ferramentas de domínio específico ou transformação (e.g. limpeza de dados) e de auditoria dos dados [Chaudhuri & Dayal, 1997] [Jarke et al., 2003] [Raman & Hellerstein, 2001] ou a distinção entre ferramentas de âmbito genérico (e.g. tratamento de diferentes tipos de anomalias nos dados) e ferramentas de âmbito restrito (e.g. correcção de nomes e moradas) [Oliveira et al., 2004]. Nesta dissertação, adoptamos a enunciação das diferentes propostas de qualidade dos dados em dois grupos principais: as propostas independentes ou protótipos de investigação e as ferramentas comerciais. Porém, abordaremos preferencialmente as propostas de investigação pertencentes ao domínio académico porque possuem, inerentemente, uma independência sob os aspectos comerciais e de *marketing*, questão indissociável na outra tipologia de ferramentas. Bem como, descrevem de forma objectiva e transparente a estrutura das arquitecturas consideradas e os modelos e técnicas adoptadas.

### 4.4.1 Protótipos de investigação

Apesar dos protótipos de investigação denotarem, regra geral, uma apetência para o tratamento alargado dos defeitos nos dados, iremos tentar circunscrever estas propostas em grupos homogéneos que procurem, predominantemente, referenciar um domínio específico. Assim, consideramos três categorias principais: auditoria dos dados, duplicação dos dados e transformações nos dados.

## Propostas de auditoria dos dados

A auditoria dos dados emprega algoritmos de mineração de dados para aferir e melhorar a qualidade dos dados. A auditoria pode ser dividida em duas tarefas: a indução da estrutura e a detecção dos desvios. A indução da estrutura corresponde a uma descrição sobre a estrutura dos dados existentes. Enquanto, a detecção dos desvios ocupa-se da verificação dos dados que denotam desvios capazes de revelarem possíveis erros e a geração das acções correctivas desses erros [Jarke et al., 2003].

A proposta de apresentada em [Marcus & Maletic, 2000] consiste na aplicação de algoritmos de mineração de dados na detecção automática de erros na base de dados do centro de investigação e desenvolvimento da marinha dos E.U.A.. O objectivo final da proposta passa pela aprovação dos métodos considerados como capazes de identificarem os valores fora dos domínios considerados e que se revelam errados. Os métodos utilizados são: a detecção estatística de valores fora dos limites, a segmentação dos dados, baseada em padrões e as regras de associação.

O protótipo apresentado em [Jarke et al., 2003] define-se como um ambiente de auditoria dos dados e pretende resolver as anomalias nos dados através de processos que envolvam métodos de aprendizagem pela máquina, capazes de induzir estruturas compreensíveis semanticamente sobre os dados existentes. Deste modo, torna-se possível classificar os valores situados fora dos limites definidos como erros potenciais. A investigação assenta em dois processos principais: a geração de dados de teste e a indução da estrutura e detecção dos desvios. A proposta pressupõe também como actividades: a análise do domínio, a selecção e adaptação do algoritmo de mineração dos dados a usar e a interactividade com o utilizador no momento da correcção. A análise de domínio consiste na identificação por um perito das propriedades estruturais dos dados. Assim, encontram-se criadas as condições para a criação de um banco de dados de teste. Baseados nestes dados de teste é possível testar a aplicação de diferentes algoritmos de mineração. Estes algoritmos são sujeitos a adaptações contínuas e são testados até que os resultados sejam considerados satisfatórios. Por fim, a ferramenta de auditoria dos dados é usada por um engenheiro de dados em vista a detecção de erros e a definição de sugestões para a obtenção de valores correctos na base de dados.

## Propostas de tratamento de duplicados

### Método da reunião e remoção

A investigação apresentada em [Hernandez & Stolfo, 1998] consiste numa proposta de identificação e eliminação de valores duplicados em conjuntos de dados. Regularmente, os métodos de identificação de duplicados assentam em algoritmos de pesquisa dos dados na busca de registos

referentes a uma mesma entidade. Esta proposta [Hernández & Stolfo, 1995, 1998] apresenta-se como um dos métodos, especializados na identificação de duplicados, mais referidos nas investigações desenvolvidas. A proposta consiste num método eficiente que reduz o número de comparações necessárias, designado por método da vizinhança ordenada. As linhas são ordenadas em função duma chave construída a partir das colunas duma tabela, na esperança que as linhas duplicadas se encontrem perto umas das outras. Em seguida, apenas as linhas que se encontram no interior duma janela são comparadas entre si de modo a detectar a possível duplicação. A janela desloca-se ao longo das linhas que fazem parte da tabela. A classificação dum par de linhas como duplicado é baseada em regras. Em vista aumentar a precisão do algoritmo, os resultados das diversas passagens (deslocamentos das janelas) são combinados através da transitividade entre todos os pares de duplicados encontrados, designando-se esta passagem como o método multi-passagem sobre a vizinhança ordenada [Oliveira et al., 2004].

Este método foi posteriormente refinado para dar resposta a situações em que novos dados são adicionados aos dados anteriormente inspeccionados e tratados (situação típica em ambientes de DW). O algoritmo original prevê que todo o conjunto de dados seja inspeccionado, mas o novo método assume uma perspectiva incremental. A postura incremental baseia-se em informações anteriores, recolhidas das execuções do método de reunião e remoção. O objectivo consiste na redução do tempo dispendido, em comparação com método original, na identificação e junção de duplicados [Hernandez & Stolfo, 1998].

**Eliminação de linhas ligeiramente duplicadas**

A proposta de [Ananthakrishna et al., 2002] consiste, também, numa aproximação visando a eliminação de valores duplicados. É expressa uma proposta que pretende evitar os problemas dos métodos de ordenação (e.g. método de reunião e remoção), como seja o elevado número de falsos positivos originados por esses métodos. Esta proposta faz uma destrinça clara entre a detecção de duplicados em SGBDs relacionais (as colunas de duas linhas têm de ser iguais) e a detecção de duplicados em operações de limpeza de dados, que apresentam questões mais subtis. Daí, a assumpção como o problema das linhas vagamente duplicadas. A investigação tem como objecto de estudo as tabelas dimensão envolvidas num esquema multidimensional dum DW. A proposta consiste num algoritmo (*Delphi*), que explora as hierarquias das tabelas dimensão, em vista a remoção de duplicados e a redução dos falsos positivos que os outros métodos provocam. Esta proposta recorre ao método apresentado em [Hernández & Stolfo, 1995], em especial, na adopção da estratégia de janela para comparar todas as linhas situadas dentro dela.

**IntelliClean**

A proposta *IntelliClean* [Lee et al., 2000a] ocupa-se, essencialmente, da eliminação de valores duplicados em conjuntos de dados. A proposta baseia-se em três etapas: o pré-processamento, o processamento e a verificação e validação humana. A primeira etapa respeita à análise e correcção das irregularidades sintácticas dos dados e à uniformização consistente de tipos de dados, formatos e abreviaturas. Na etapa do processamento estabelecem-se regras de limpeza dos dados pré-processados. As regras consideradas são classificadas em quatro tipos. As regras de identificação de duplicados especificam quando duas linhas são consideradas duplicadas. As regras de fusão e remoção indicam quando é que duas linhas devem ser fundidas. As regras de actualização indicam como as linhas são actualizadas. As regras de alerta definem as condições de aviso aos utilizadores. Por último, a etapa de verificação e validação compreende a intervenção humana em vista aferir a eficácia das operações realizadas nos dados e eventualmente proceder à sua correcção, como seja a decisão de linhas para as quais não foram definidas regras de fusão e remoção ou corrigir situações em que as linhas foram consideradas erradamente como duplicadas [Müller & Freytag, 2002] [Lee et al., 2000a] [Low et al., 2001].

## Propostas de transformação dos dados

**Sistema AJAX**

O sistema AJAX [Galhardas et al., 2000a, 2000b] é uma proposta assente numa arquitectura flexível e extensível que procura separar o nível lógico e físico. O nível lógico especifica as operações de limpeza dos dados a realizar e o nível físico, trata dos aspectos relacionados com a implementação. O principal objectivo consiste em facilitar a transformação dos dados existentes numa ou mais fontes de dados num determinado esquema alvo e executar, simultaneamente, um conjunto de acções de remoção de defeitos nos dados. A proposta considera cinco tipos de operações de transformação sobre os dados: o mapeamento, a vista, a correspondência, a segmentação e a fusão [Galhardas et al., 2000] [Oliveira et al., 2004] [Müller & Freytag, 2002].

O mapeamento consiste na uniformização entre diferentes formatos dos dados (e.g. formato das datas) ou simplesmente na fundição ou divisão de atributos sob um formato predefinido. A transformação de vista corresponde ao operador vista de SQL, permitindo especificar as uniões e junções em SQL e verificar violações de integridade referencial dos dados. A correspondência respeita à identificação de pares de linhas que, com grande probabilidade, correspondem à mesma entidade ou objecto. As linhas são comparadas usando uma ou várias colunas definidas como critérios para a duplicação de valores dos dados. O terceiro tipo de transformação, a segmentação, baseia-se nos resultados da operação de transformação precedente e agrupa os pares de linhas

que aparentam um elevado grau de semelhança. Por último, a operação de fusão consiste na aplicação sobre cada segmento de linhas, obtido na operação anterior, com objectivo de eliminar os valores duplicados [Galhardas et al., 2000] [Müller & Freytag, 2002]. A investigação trata o assunto da eliminação dos valores duplicados como o problema de identificação do objecto.

O processo de remoção de anomalias dos dados é modelado como um grafo dirigido das operações de transformação sobre o fluxo de dados. A arquitectura alberga ainda um mecanismo de linhagem dos dados que possibilita aos utilizadores do sistema a análise da proveniência dos dados e deste modo, contribuir para a correcção dos valores que deram origem a deficiências nos dados [Müller & Freytag, 2002].

**Potter's Wheel**

A proposta *Potter's Wheel* [Raman & Hellerstein, 2001] disponibiliza um sistema de limpeza dos dados que integra num único interface, simultaneamente, a detecção das discrepâncias dos dados e as operações de transformação a executar sobre os mesmos. Este sistema permite aos utilizadores, gradualmente, procederem à composição e análise das operações de transformação sob um interface gráfico e intuitivo, do tipo folha de cálculo [Low et al., 2001]. O utilizador define os resultados a atingir sobre uma amostra de dados e automaticamente são inferidas as expressões que descrevem o domínio especificado. Deste modo, os utilizadores não necessitam de especificar as expressões antecipadamente [Raman & Hellerstein, 2001] [Müller & Freytag, 2002]. As transformações deste sistema são muito similares aos operadores de reestruturação oferecidos pelo *SchemaSQL*. Contudo, a solução aqui apresentada visa a facilidade na especificação e a aplicação incremental das operações de transformação e não apenas mostrar a potência dessas transformações [Raman & Hellerstein, 2001].

Os tipos de discrepâncias nos dados considerados neste sistema são: a discrepância estrutural, a discrepância esquemática e a violação do domínio. A discrepância estrutural revela-se como consequência entre diferenças ao nível do formato dos campos (e.g. 31/Out/2005 e 2005/10/31). Quanto à discrepância esquemática, resulta de uma deficiente estratégia de integração dos dados provenientes de múltiplas fontes. Por último, a detecção da violação do domínio pode ser melhorada pela aplicação de algoritmos específicos do domínio. Esta discrepância pode assumir duas formas distintas: envolvendo uma única linha, quando o valor de uma coluna viola as restrições referentes ao seu domínio e envolvendo várias linhas, ou seja, quando o valor de uma ou mais linhas violam uma restrição, ainda que, individualmente, cada linha esteja correcta (e.g. violação de dependência funcional) [Oliveira et al., 2004].

**ARKTOS**

O ARKTOS é uma proposta que permite a modelação e execução do processo de ETL, em ambientes de DW, baseado num conjunto de primitivas para a realização de tarefas de transformação e limpeza comuns nestes processos [Vassiliadis et al., 2001]. As tarefas de limpeza dos dados estão incluídas no processo de ETL, que envolve igualmente a extracção cirúrgica de dados relevantes das fontes, a sua transformação numa uniformização definida e por fim, a execução do carregamento num DW. A proposta designa as operações do processo de ETL como actividades, constituindo-se estas como unidades atómicas de trabalho. Dado que a finalidade de uma actividade é efectuar uma acção no processo de ETL, então cada actividade encontra-se conectada a tabelas de um ou mais conjuntos de dados. A cada actividade encontra-se também associada um tipo de erro particular e uma política da acção a tomar. A lógica subjacente a cada actividade é descrita através de uma instrução SQL [Vassiliadis et al., 2001] [Müller & Freytag, 2002].

O ARKTOS procura fornecer funcionalidades que colmatem as dificuldades comuns de complexidade e manuseamento inerentes a um cenário de ETL. Assim, são oferecidos dois modos de definição das actividades: o modo gráfico e o modo declarativo [Vassiliadis et al., 2001]. O primeiro, fornece um conjunto de actividades que correspondem às operações mais usuais de transformação e limpeza, de modo a suportar o processo de ETL e que visam tratar as seguintes violações: chave primária, integridade referencial, unicidade, valor nulo, formato e domínio. O segundo modo potencia o recurso a duas linguagens declarativas: a XADL, orientada para uma fácil e compreensível descrição do cenário e a SADL, destinada a suportar a definição declarativa do cenário de ETL, num estilo de SQL [Vassiliadis et al., 2001]. O sucesso da limpeza dos dados pode ser medido para cada actividade através da execução do comando SQL que faça a contagem das linhas que correspondem ou violam as diversas regras [Müller & Freytag, 2002].

### 4.4.2 Ferramentas comerciais

Relativamente às propostas apresentadas no domínio comercial podemos assistir a uma panóplia de ferramentas que abarcam, maioritariamente, as várias actividades inerentes à gestão dos dados em SDWs, em especial, aquelas relativas ao tratamento de imperfeições verificadas nos dados. O critério de escolha adoptado baseou-se nas ferramentas comummente referenciadas em publicações académicas e que denotam uma apetência para ambientes de DW. Inicialmente, as ferramentas abordavam, regra geral, o tratamento específico de uma acção ou de um conjunto diminuto de acções sobre os dados (e.g. a decomposição ou estandardização dos dados). Esta situação configurava uma necessária versatilidade na integração entre ferramentas capazes de garantir a complementaridade das actividades necessárias.

Actualmente, assiste-se ao desenvolvimento de ferramentas que procuram garantir a cobertura de todas actividades relativas ao *back-end* dos SDWs. Tendencialmente, os produtores de *software* têm vindo a aperfeiçoar os seus produtos como resposta às exigências actuais de uma gestão mais abrangente dos dados. O reconhecimento da importância da gestão dos dados em SDWs tem levado ao aparecimento de ferramentas que, conceptualmente, respeitam metodologias comprovadas, como seja o caso da TDQM e o ciclo PDCA. Todavia, estas soluções abrangentes de gestão dos dados revelam, por vezes, limitações de operacionalidade com o produto e apresentam-se como marcas proprietárias nos formatos de metadados, o que pode inviabilizar a conciliação de outras ferramentas [Rahm & Do, 2000]. A enunciação das ferramentas encontra-se disposta de acordo com o aspecto específico a tratar no fluxo circulatório dos dados no SDW: análise dos dados, transformações e limpeza dos dados e ferramentas de ETL.

## Análise dos dados

Esta classe de ferramentas ocupa-se da identificação de erros e inconsistências nos dados. As ferramentas nesta área procuram responder a questões comuns aquando da análise dos dados. Nomeadamente, sobre os dados e as fontes sobre a validade, a completude e o cumprimento das regras de negócio.

A *WizWhy & WizRule*, da *WizSoft, Inc.* [9], são duas aplicações que procuram inferir sobre os relacionamentos e regras entre as colunas e os seus valores. As regras permitem o estabelecimento de predições e revelam os padrões ocorridos nos dados. A *WizWhy* é uma aplicação de mineração de dados, que baseada em regras e padrões, permite estabelecer predições sobre os valores em casos futuros. O *WizRule* analisa as bases de dados e revela três tipos de regras: fórmulas matemáticas, estruturas condicionais e baseadas em dicionários. A ferramenta usa as regras e aponta os desvios relativamente a estas regras como erros nos dados [9] [8] [Rahm & Do, 2000].

A *AXIO* [8] disponibiliza um ambiente de *profiling* que descreve o conteúdo, a estrutura e as complexas estruturas de dados das bases de dados. A aplicação executa as acções de *profiling* em três dimensões: as colunas, as dependências e a redundância de valores. Os resultados obtidos pela execução da ferramenta permanecem num repositório e constituem-se como metadados sobre os dados avaliados.

## Correcção e estandardização de dados

As ferramentas que fazem parte desta categoria são usualmente específicas no domínio de irregularidades a tratar. Geralmente, as técnicas consideradas consistem na extracção e transformação dos dados em elementos elementares estandardizados e individualizados (e.g. nomes e moradas).

O sistema proposto pela *Trillium Software* [10] disponibiliza duas aplicações: *Trillium Software Discovery* e *Trillium Software System*. A primeira consiste numa aplicação de análise e *profiling* dos dados, com o objectivo de revelar o conteúdo real dos dados. A ferramenta comporta também a possibilidade de monitorização dos dados, a validação dos dados com as regras do negócio e a análise de tendências. A segunda aplicação compreende as acções de limpeza e transformação dos dados através de quatro etapas: a investigação (informações sobre os registos actuais); a estandardização (representação dos dados de modo consistente); o enriquecimento (complementação dos dados existentes) e a ligação (identificação dos relacionamentos entre as linhas). A integração das duas aplicações pretende seguir uma linha orientadora da metodologia TDQM.

O *software Athanor* [8] possibilita uma compreensiva solução de tratamento da qualidade dos dados assente em cinco etapas: a auditoria dos dados, para medir o nível de qualidade dos dados e a natureza dos problemas com os dados; a aplicação de regras e objectivos, através de componentes disponibilizados pela ferramenta; o estabelecimento de planos de qualidade dos dados para a configuração de um conjunto de regras de estandardização; a execução dos planos em conjuntos de dados e as listagens das actividades desenvolvidas pelas acções executadas e que permitem uma monitorização da qualidade dos dados.

## Duplicação de valores

A duplicação de dados é um dos desafios mais comuns e difíceis de enfrentar em ambientes de DW. A reunião de dados provenientes de diferentes localizações e respeitando esquemas e formatos diferenciados dificulta ainda mais este processo. Assim, esta tarefa deve ser executada após as etapas comuns de transformação dos dados e no momento da reunião dos mesmos.

A ferramenta *MatchIT* [12], da *helpIT Systems Limited*, apresenta um alto nível de interactividade com os utilizadores e permite a estes a especificação de critérios de correspondência entre linhas, através da combinação de funções aplicáveis às colunas. A confrontação entre linhas é apenas realizada sobre aquelas que respeitam o mesmo critério de correspondência. Deste modo, realizam-se diferentes pesquisas para diferentes critérios. A enunciação dos critérios pode ser aperfeiçoada por uma matriz de importância, que atribui diferentes pesos aos critérios a ponderar. A confrontação é realizada coluna a coluna das linhas comparadas. As linhas identificadas como semelhantes são agrupadas num local, constituindo segmentos de linhas [7] [Rahm & Do, 2000].

O *Group1 Software* [13] disponibiliza uma ferramenta, *Merge/Purge Plus*, que permite inicialmente, efectuar as operações de limpeza dos dados em nomes e moradas e posteriormente, realizar operações de correspondência entre linhas. A ferramenta fornece diferentes opções de correspondên-

cia entre linhas, desde ligeiramente duplicadas até à duplicação integral das linhas. Após a definição da linha representativa, é possível a remoção das linhas não desejadas [Neely, 1998] [7] [Rahm & Do, 2000].

## Ferramentas de ETL

Algumas ferramentas comerciais suportam, de modo compreensivo, os processos de ETL em SDWs. Geralmente, recorrem a um repositório assente num SGBD que efectua a manutenção dos metadados, de modo integrado e uniforme, sobre as fontes de dados, esquemas alvo, mapeamentos e processos envolvidos. Os esquemas e os dados são extraídos do SO por meios de ligação estandardizados. As operações de transformação e limpeza dos dados são realizadas de maneira acessível e permitem a interacção do utilizador. Normalmente, na etapa de mapeamento recorre-se a uma linguagem de regras proprietária e a uma biblioteca de funções de conversão predefinidas (e.g. formatos dos dados). Algumas propostas permitem a possibilidade de incorporar ferramentas externas capazes de realizar um tratamento específico mais adequado (e.g. limpeza de nomes e endereços e eliminação de duplicados) [Rahm & Do, 2000].

O pacote *dfPowerStudio*, disponibilizado pela *Dataflux* [14], pretende abarcar as actividades do processo de ETL. A ferramenta procura seguir conceptualmente o ciclo PDCA e nesse propósito estabelece a gestão dos dados numa sequência de cinco grandes actividades: o *profiling*, a qualidade, a integração, o enriquecimento e a monitorização dos dados. Deste modo, a gestão dos dados fornece as funcionalidades necessárias para a construção de um repositório de dados consistente, correcto e fiável.

A *NCR* disponibiliza duas ferramentas, *Teradata Warehouse* e *Teradata Warehouse Miner*, capazes de assegurarem de modo integrado e consistente as tarefas de *Back-end* que envolvem os SDWs. A *Teradata Warehouse* permite que os processos envolventes à qualidade dos dados sejam garantidos dentro da arquitectura dum SDW e compreende como principais componentes: o motor de regras, o *profiling* e auditoria dos dados e as operações de transformação e limpeza. A primeira componente, o motor de regras, é o local que define os processos de qualidade dos dados e as métricas usadas para determinar essa qualidade. O *profiling* e auditoria dos dados permitem a aferição quer dos dados existentes nas fontes, quer dos dados localizados no repositório do DW. As operações de transformação e limpeza decorrem durante as fases de propagação dos dados para a ARD e para o repositório do DW [Gonzales, 2003].

A ferramenta *Teradata Warehouse Miner* cobre uma área alargada das tarefas de *profiling*, auditoria, limpeza e transformação dos dados. O *profiling* dos dados compreende o entendimento sobre

os dados existentes no SO, em especial: a correcção, a consistência, a completude e a integridade. Para isso, são fornecidas as seguintes funções: análise de valores, frequências, análises estatísticas, histogramas e árvores de decisão. A auditoria dos dados considera um policiamento ou monitorização dos dados inseridos nas fontes, baseada no *profiling* dos dados obtido e nos níveis de qualidade especificados (e.g. frequência, histograma, etc.). As operações de limpeza permitem a alteração do código SQL gerado pela aplicação de modo a realizar as tarefas pretendidas. A ferramenta disponibiliza várias funções que constituem as operações de transformação mais comuns: a integração, a agregação e a estandardização [Gonzales, 2003].

Uma outra ferramenta, recente no mercado, procura disponibilizar uma solução potente e flexível, capaz de efectuar uma eficiente gestão dos dados em SDWs [Iwaysoftware, 2004]. Neste sentido, é dotada de quatro características: acesso directo ao SO, soluções de ETL robustas e rápidas, gestão compreensiva dos metadados e análise dos recursos envolvidos. A primeira, o acesso directo ao SO, pretende salientar a facilidade dos adaptadores no acesso a fontes heterogéneas. A segunda respeita a uma gestão do processo de ETL compreensiva e potente, que assenta num conjunto de ferramentas que simplificam a criação, manutenção e expansão do DW. A facilidade de utilização assenta na interactividade na condução das operações pelo utilizador. Em seguida, o processo de ETL incorpora as capacidades de *Change Data Capture* (CDC). Esta capacidade responde perante as necessidades de gestão incremental dos dados e redução da janela de oportunidade disponível, ou seja, apenas os dados inseridos, actualizados e removidos são considerados em futuras migrações dos dados provenientes das fontes e a carregar no repositório de DW. Por último, todo o sistema encontra-se suportado numa plataforma de metadados que possibilita uma gestão integrada, compreensiva e aberta dos mesmos [Iwaysoftware, 2004].

## 4.5 Administração dos dados

Normalmente em ambientes de DW, a atribuição de padrões de qualidade aos dados é concretizada durante as tarefas do processo de ETL, que ocorre maioritariamente na ARD. Todavia, as exigências actuais, tanto do ponto de vista económico-financeiro, como do ponto de vista da eficácia na logística processual, não se compadecem com ferramentas e técnicas agindo de modo independente e não obedecendo a um plano de qualidade integrador das acções a tomar em termos da gestão dos dados em SDWs. As imposições ao nível da qualidade dos dados em SDWs devem conduzir a propostas que prevejam de modo metódico a compreensão dos processos envolvidos ao longo do sistema circulatório dos dados e dos metadados associados. Assim, pretende-se garantir a manutenção de dados com elevado grau de confiança de forma a permitir atingir dois dos objectivos basilares dum SDW: a gestão coerente dos dados organizacionais e a criação

de vantagens estratégicas face à concorrência pela oportuna utilização dos dados. As iniciativas que visam a melhoria a qualidade dos dados nos SDWs devem encontrar-se sustentadas e formalmente constituídas numa área da estrutura organizativa. Esta possibilidade pressupõe o reconhecimento pela organização dos problemas nos dados como um assunto que cobre toda a organização e cujas responsabilidades devem ser assumidas por todos e consequentemente, considerar a qualidade dos dados como fazendo parte dos processos de negócio. Neste sentido, as responsabilidades com os dados devem deixar de estar sobre alçada das tecnologias de informação [Adelman et al., 2005]. Em [Redman, 1995] refere-se os dados como uma área de gestão autónoma e que pressupõe o envolvimento da cúpula organizacional [6].

A abordagem seguida na constituição desta área funcional procura responder a algumas orientações sugeridas no âmbito do projecto DWQ [Jarke & Vassiliou, 1997] [Vassiliadis, 2000]. Em especial, tendo presente uma estrutura, que trate os assuntos da qualidade dos dados, composta por: política, gestão, sistema, controlo e segurança. A política da qualidade dos dados transmite a intenção e a direcção a seguir, por parte da organização, relativamente às preocupações a nível da qualidade dos resultados produzidos. A gestão da qualidade dos dados compreende a responsabilidade pela execução das tarefas funcionais em vista a implementação da política de qualidade dos dados. O sistema de qualidade dos dados congrega a estrutura organizacional, as responsabilidades, os procedimentos, os processos e os recursos para a implementação da gestão da qualidade dos dados. O controlo da qualidade é o conjunto de actividades e técnicas operacionais usadas em vista a obtenção da qualidade desejada nas informações produzidas. Por último, a segurança da qualidade inclui todos os planeamentos e acções necessárias de forma a assegurar que a conformidade das informações obtidas satisfaz os requisitos de qualidade impostos.

### 4.5.1 Motivos

A possibilidade de criação de uma área funcional, responsável pela gestão dos dados organizacionais, não deve ser entendida como o esvaziar de funções incumbentes aos administradores do SDW. Antes pelo contrário, a impossibilidade dos administradores do DW intervirem no SO conduz, futuramente, ao agravamento da natureza das tarefas inerentes à ARD. Especialmente, quando consideradas as últimas tendências dos dados manipulados pelos SDWs e que apontam para os SDWs de segunda geração [Inmon, 2006b]. A compreensão de dados estruturados e de dados sem estrutura, provenientes de mensagens de correio electrónico, de voz e de imagens afectará, certamente, as tradicionais tarefas de tratamento dos dados. O incumprimento dos orçamentos previstos será, ainda mais, uma realidade caso não sejam antecipadas acções que promovam a qualidade nos dados. A clareza do caminho a seguir mostra-se ainda mais visível se considerar-

mos que as operações ocorridas na ARD correspondem a 2/3 do orçamento total previsto para a implementação dum SDW. Estas operações transformam e melhoram os dados organizacionais para o propósito da decisão, mas mantêm irregulares os dados existentes no SO, situação que configura um desperdício de recursos que podem ser canalizados noutras actividades.

### 4.5.2 Objectivos

A área funcional dos dados deve ter presente um desígnio que a oriente no cumprimento dos objectivos traçados. O desígnio deverá assentar na prevenção da presença de defeitos nos dados, em geral, na organização e em particular, nos SDWs. Em [Adelman et al., 2005] são descritos como objectivos subjacentes à área de administração dos dados a:

- A responsabilidade pela qualidade dos dados perante a organização.
- A coordenação e execução das actividades de inspecção e reparação dos defeitos.
- A administração dos metadados referentes aos dados e processos envolventes.
- A prevenção na captação de dados defeituosos nas fontes de dados.

### 4.5.3 Actividades

A área dos dados deve prever o cumprimento de um conjunto de actividades desenvolvidas pelos diversos responsáveis dos dados. Esta área deve ser preenchida por administradores dos dados, administradores dos metadados e guardiães dos dados. Os últimos devem prevenir a propagação dos defeitos nos dados pela organização, em especial, pelos agentes de decisão. Os administradores dos metadados devem realizar a gestão dos metadados, em particular, na difusão de informações relevantes sobre os dados disponibilizados. Os administradores dos dados devem ser responsáveis pelo modelo de dados organizacional, em impulsionar a execução de uma política dos dados que vise a prevenção em detrimento da inspecção e correcção dos defeitos nos dados, em designar os standards e linhas guia na captura dos dados inseridos no sistema. Algumas das actividades que devem ser desenvolvidas no âmbito da administração dos dados devem incluir:

- A formação dos intervenientes, em particular, dos agentes de decisão.
- A identificação e estabelecimento de prioridades dos dados críticos.
- O alinhamento da estratégia dos dados aos objectivos organizacionais.
- A implementação de uma política de qualidade dos dados.

- A prevenção dos defeitos nos dados (e.g. formulários mais intuitivos, estandardização de valores, normalização de padrões e linhas de orientação na introdução de dados).
- A distribuição de métricas estrategicamente colocadas para aferição do sucesso do DW.
- A introdução de incentivos para a prevenção da captura de dados irregulares.
- A monitorização e melhoramento contínuo dos processos geradores de defeitos.
- A formação dos diversos intervenientes para estarem alerta sobre a qualidade dos dados, devendo estes informar os responsáveis da ocorrência de defeitos nos dados.
- A execução de auditoria aos dados para determinar as causas dos defeitos nos dados.
- A monitorização dos melhoramentos da qualidade dos dados.
- A promoção da mudança cultural da organização.
- A manutenção de metadados referentes aos processos, dados e objectos dos SDWs.

### 4.5.4 Prevenção dos problemas nos dados

As ferramentas tradicionais, apesar de necessárias, pecam por não fornecerem uma aproximação sistemática sobre a administração dos dados nos SDWs [Shankaranarayan, 2005]. Estes métodos sofrem de duas falhas. Primeira, pressupõem a avaliação da qualidade dos dados de modo imparcial à contextualização da utilização dos mesmos. Segunda, estes métodos não permitem a avaliação da qualidade dos dados numa etapa específica usando as medidas associadas às etapas precedentes. Ainda segundo Shankaranarayan, os dados em SDWs são processados por um conjunto de etapas sequenciais e a qualidade dos dados associada a uma etapa depende directamente dos níveis de qualidade nas etapas precedentes. Esta problemática configura a linhagem dos dados como uma das preocupações mais prementes no fluxo circulatório dos dados em SDWs.

Subjacente à aplicação dos métodos e técnicas tradicionais encontra-se presente, regra geral, uma estratégia de reparação das irregularidades dos dados. Nesta dissertação, pretendemos seguir uma linha condizente com a prevenção da ocorrência de anomalias nos dados em detrimento da opção tradicional de detecção e reparação dos problemas verificados. Deste modo, tenta-se dar provimento ao conceito de PI e às técnicas e modelos usados na gestão da qualidade dos produtos convencionais. A este respeito, em [Amaral et al., 2002] consideram-se as propostas tradicionais, de identificação e reparação de defeitos nos dados, como soluções de curto prazo porque os processos causadores dessas deficiências se mantêm inalterados. Em [Kimball & Caserta, 2004] encaram-se as fontes dos dados, como o local em que, maioritariamente, a resolução

dos problemas e a melhoria dos dados deve ser tratada. É igualmente referido que o tratamento dos problemas nos dados a montante mostra-se como a única estratégia defensiva a tomar no sentido da melhoria dos dados nas organizações.

Neste contexto, torna-se necessário e urgente substituir a estratégia tradicionalmente adoptada, de inspecção e reparação, por estratégia que vise a prevenção da ocorrência de deficiências nos dados. A aplicação desta estratégia impõe o envolvimento dos responsáveis organizativos nestas questões, bem como em assumir os problemas nos dados como uma assunto organizacional. Nomeadamente, a compreensão que a tomada de melhores decisões, assentes nos SDWs, é fortemente condicionada pela eficiência na gestão dos dados. Esta é a condição essencial para a mudança de filosofia e cultura da organização, no que concerne à qualidade dos dados. Assim, pretende-se dar aval às investigações mais recentes que apontam no sentido da impossibilidade duma efectiva gestão da qualidade dos dados em SDWs sem o entendimento, pelos responsáveis organizacionais, que este é um assunto transversal a toda a organização [Kimball & Caserta, 2004] [English, 2004] [Shankaranarayan, 2005] [Gonzales, 2004].

Esta nova postura no confronto com os problemas dos dados acarreta naturalmente consequências, em especial, o facto do âmbito da acção dos processos de melhoria da qualidade dos dados, tendencialmente, extravasar o habitual domínio dos SDWs. Contudo, é justificada em primeiro lugar porque o investimento em recursos (financeiros, materiais e humanos) não se deve compadecer com a continuação de imperfeições nos dados das fontes. Em segundo lugar, uma vez concretizada a melhoria dos dados revela-se uma menos valia a permanência de dados errados nas fontes, que inviabilizem o progresso de outras aplicações informáticas. Por último, as exigências quotidianas de mais e melhor informação determinam a necessidade em tratar estas questões o mais próximo da captura dos dados, de forma mais célere e actuando preventivamente (e.g. a recolha atempada e completa de informações vitais, sobre o estado de saúde dos pacientes, pode possibilitar a elaboração de melhores árvores de decisão em aplicações de mineração dos dados).

Ora, estas contingências configuram um deslocamento de algumas das acções de tratamento dos dados para uma fase a montante do fluxo circulatório dos dados em SDWs. Este cenário parece revelar uma propensão para o rompimento entre a fronteira que delimita o domínio dos SDWs e o SO. Todavia, não se pretende aqui estabelecer um novo tipo de arquitectura para os SDWs, mas salientar, por um lado, a necessidade desta problemática ser tratada numa perspectiva global e a um patamar superior da hierarquia organizativa. Por outro lado, estabelecer uma separação clara entre os procedimentos de limpeza e tratamento dos dados que devem ser executados no âmbito do SO e os que são exclusivamente tratados no domínio do DW.

# Capítulo 5

# A Aferição da Qualidade dos dados em SDWs

O exercício da tomada de decisão, no contexto de ambientes altamente dinâmicos e imprevisíveis, implica necessariamente a propensão para o controlo de um conjunto de variáveis pelos decisores [Shankaranarayan et al., 2003]. Os SDWs assumem particular preponderância como uma plataforma tecnologicamente capaz de minimizar o risco e de potenciar a eficácia decisória. Os dados existentes num SDW suportam tecnicamente a decisão e por isso, a sua qualidade é determinante como forma de fomentar a relação de confiança com os decisores. A gestão da qualidade dos dados deverá seguir como orientação principal a adequação dos dados ao uso dos decisores, satisfazendo ou superando as suas necessidades e realçando as características mais caras a estes. Assim, a importância em medir os desempenhos dos dados, nos mais variados quadrantes, parece ser uma forma efectiva na conquista de elevados padrões de qualidade destes, nomeadamente, no reconhecimento intrínseco dos seus pontos fortes e fracos. O conhecimento sobre estes aspectos permite potenciar virtudes e melhorar as deficiências existentes porque clarifica os objectivos a atingir pela aplicação duma estratégia que vise a melhoria dos dados [Bobrowski et al., 1998]. Assim, pretende-se dar provimento ao princípio: se não medir, não pode gerir.

"O funcionamento de um DW sem recurso a métricas é semelhante a navegar um navio sem um compasso, um mapa ou uma bússola." [Adelman, 2002].

As métricas permitem aferir a capacidade de funcionamento, e em consequência, o grau de sucesso atingindo num SDW. Para tal, podem e devem ser constituídas métricas sobre os diversos componentes, processos e outros recursos do DW, como sejam [Adelman, 2002]:

- Métricas sobre os níveis de utilização do DW.

- Métricas que revelem o desempenho na resposta aos utilizadores.
- Métricas indicadoras da disponibilidade do sistema.
- Métricas sobre os níveis de satisfação dos utilizadores e que compararem as expectativas iniciais dos utilizadores e os resultados atingidos.
- Métricas de qualidade dos dados, que promovam o melhoramento contínuo.
- Métricas sobre os dados dormentes.
- Métricas sobre a utilização de ferramentas fornecidas pelo DW.
- Métricas dos custos e dos benefícios.

Neste capítulo iremos centrar a nossa atenção em métricas que actuem nos SDWs para avaliar a qualidade dos dados disponibilizados. Estas métricas devem ser capazes de fornecer, sempre que possível, informações objectivas sobre os dados. A natureza multidimensional inerente à qualidade dos dados e a dificuldade associada à definição do conceito possibilitam divergentes perspectivas na avaliação destes. A divergência de opiniões resulta da ponderação atribuída pelos consumidores face ao desempenho conseguido nas diversas dimensões consideradas nos dados. Assumindo a qualidade dos dados como um conceito multidimensional e variável no tempo, a determinação do grau de qualidade dos dados expostos pelos DWs, resulta da medição conseguida em cada uma das dimensões. Neste cenário, a associação de meios de medida às dimensões, proporciona informações sobre os níveis de cumprimento em cada uma das características dos dados e logicamente, decidir sobre o sucesso do sistema.

## 5.1 Definição de métricas

Contextualizando ao âmbito dos dados, devemos construir métricas capazes de sentir o pulsar da qualidade dos dados constantes num DW. A utilidade dos dados disponibilizados encontra-se relacionada com os processos de medição da qualidade. Por sua vez, a qualidade dos dados apresentada aos utilizadores dum DW resulta de um conjunto de factores, nomeadamente, da qualidade dos processos, dos componentes e dos dados [Naumann & Rolker, 2000]. Em [Basili et al., 1994] descrevem-se os princípios norteadores da construção de métricas que avaliam o desempenho conseguido nos projectos de *software*, mas possíveis de adoptar às investigações do domínio dos SDWs [Vassiliadis, 2000] [Calero et al., 2001] [Serrano et al., 2002] [Amaral, 2003]. Assim, as medidas de avaliação são percepcionadas como mecanismos para a criação da memória da organização. Estes mecanismos permitem a concepção de planos mais realistas, o fornecimento de valores objectivos que permitam a adopção de técnicas de refinamento, o acompanhamento da

evolução dos processos ou objectos e a detecção dos processos ou objectos mais fortes ou mais fracos (e.g. a tabela dimensão aluno apresenta uma taxa de ausência dos dados de 75% na coluna telefone). Em [Cantone & Donzelli, 1998, 1999], seguidores da linha de pensamento de Basilli et al., as medidas são adoptadas para descreverem, compararem, planearem, monitorizarem e avaliarem os processos, produtos e recursos. Em [Calero et al., 2001] as medidas obtidas são usadas para indicar os processos que têm de ser melhorados, apontar os ganhos e perdas, comparar os objectivos com o desempenho actual, fornecer as informações para as equipas de avaliação e finalmente gerir com base em factos em detrimento de meras suposições subjectivas.

Para assegurar a qualidade dos dados não é suficiente socorrer ao critério abstracto "qualidade", porque os consumidores finais e outros intervenientes nos SDWs têm diferentes percepções do mesmo conceito. À luz da metodologia TQM, deve-se fazer uso da medição de critérios concretos que avaliem a qualidade para evitar o emprego de "argumentos de estilo" [Calero et al., 2001]. Portanto, o objectivo central no processo de desenvolvimento de métricas avaliadoras da qualidade dos dados e do desempenho do sistema deverá orientar a substituição ponderada de noções de qualidade intuitivas por medidas formais e quantitativas de forma a reduzir a subjectividade e as referências no processo de avaliação.

Assim, as métricas podem definir-se como instrumentos específicos possíveis de serem usados na medição de um determinado factor de qualidade de modo consistente e objectivo [Bouzeghoub & Peralta, 2004] [Calero et al., 2001]. As métricas podem corresponder a medidas de avaliação sobre os dados e elementos de processos envolventes (e.g. entrada, transformação, tratamento e saída) [Smith, 2004a]. O desenvolvimento de métricas que actuem sobre a qualidade dos dados de modo estandardizado possibilita uma maneira estável e consistente de expressar os resultados sobre as dimensões [Wood, 2002]. A manutenção de um histórico sobre as métricas adoptadas e os valores por elas registados permite traçar objectivos de melhoramento sobre as métricas, efectuar medidas de evolução e comparações de *benchmark*, internas e externas, sobre os níveis obtidos. As organizações anseiam pela obtenção de uma medida, que se constitua num índice sobre a qualidade dos dados e que resulte da agregação de valores sobre as diversas métricas [Pipino et al., 2002]. Este índice sugere um valor indicativo sobre a qualidade dos dados e permite efectuar comparações com as suas congéneres de modo mais preciso, efectuar melhores análises e promover a melhoria contínua dos processos e dados.

O processo de definição de métricas rege-se por dois princípios norteadores: a relevância e a validade [Smith, 2004b]. Em [Loshin, 2005] é definido que as métricas devem ser: definidas objectivamente, possíveis de medir, relevantes para o negócio e controláveis. A construção de cada mé-

trica deve pressupor um objectivo concreto a alcançar. Os objectivos traçados pelas organizações são o ponto de partida na definição de métricas e apenas estas podem ser consideradas relevantes. A validade das métricas mostra-se mais difícil de concretizar porque a obtenção de valores erróneos pode revelar-se fatal nas organizações. Em especial, se considerarmos alguns sistemas de métricas mais complexos e resultantes de métricas a montante e servindo de entrada para outras a jusante. É igualmente crucial a gestão estratégica do ciclo de vida das métricas. O ciclo de vida deve definir os procedimentos de análise, desenho, desenvolvimento, monitorização, ajustamento e a eventual supressão das métricas. Durante este processo deve ser mantida e atestada a relevância e a validade das métricas [Smith, 2004b].

## 5.2 Classificação das métricas

Para um enriquecimento da avaliação da qualidade dos dados constantes no sistema, as organizações devem operar com instrumentos de medida sobre os objectos do DW (componentes, processos e dados). As características inerentes ao DW implicam a observação por diferentes enquadramentos classificativos das métricas: o critério utilizado na determinação das métricas, o objecto de aplicação e o método de obtenção. Relativamente ao critério utilizado na determinação das métricas, podemos observar: as métricas subjectivas e as objectivas. A primeira categoria compreende as apreciações qualitativas efectuadas por parte dos consumidores finais e dos diversos intervenientes. As percepções subjectivas reflectem as necessidades e experiências dos diferentes intervenientes e são geralmente veiculadas por respostas a questionários, capazes de aferir sobre as expectativas dos vários consumidores e a qualidade dos dados fornecidos e que auxiliam o cumprimento das tarefas em mãos [Helfert & Radon, 2000] [Pipino et al., 2002]. A natureza de algumas dimensões exige uma avaliação meramente subjectiva, como sejam o caso das dimensões interpretação dos dados e a representação concisa destes [Cappiello et al., 2004]. Ainda que, esta questão deva ser amenizada pela introdução de outras observações mais objectivas (e.g. a disponibilização de metadados e sistemas de ajuda aos consumidores finais).

As métricas objectivas representam as medidas sobre as dimensões dos dados capazes de serem claramente quantificadas e obtidas através de regras claras e bem definidas. A associação destas métricas a valores quantificáveis possibilita efectuar comparações posteriores de modo consistente. Os valores observáveis devem ser sempre os mesmos, independentemente do instante, condições ou indivíduo que os determina [Abreu, 1992]. Esta classe de métricas pode ser distinguida em dois tipos: independentes da tarefa e dependentes da tarefa. As métricas independentes da tarefa reflectem o estado dos dados sem o conhecimento contextual da sua aplicação e por isso, podem ser aplicadas em qualquer sistema de dados. Geralmente, a captação dos valores destas

métricas é conseguida por mecanismos automáticos, que disponibilizam um conjunto alargado de informações estatísticas relevantes sobre os dados. Enquanto que as métricas dependentes da tarefa são desenvolvidas com base num contexto específico, necessitam do conhecimento da aplicação e formato dos dados e incluem quer imposições internas (e.g. organização) quer imposições externas (e.g. governamentais) [Pipino et al., 2002].

Quanto ao objecto de aplicação, as métricas podem ser categorizadas em: métricas de produto e métricas de processo. A primeira categoria corresponde às métricas aplicáveis à informação fornecida (e.g. as que avaliam as dimensões exactidão e completude dos dados). Enquanto que as métricas de processo referem-se aos meios de construção e desenvolvimento da informação divulgada (e.g. duração dos processos de transformação dos dados) [Abreu, 1992].

No que respeita ao método de obtenção, as métricas podem ser categorizadas em: métricas simples e métricas compostas. As métricas simples ou elementares correspondem a métricas directas e singulares sobre um único atributo não resultando de combinações com outras métricas. Enquanto que as métricas compostas são mais complexas porque resultam de valores indirectos, derivados ou escalonamentos hierárquicos. Por este motivo, estas métricas exigem maiores cuidados na sua construção e devem ser claras quanto à relevância e validade patenteadas. Estas métricas podem resultar de médias compostas ou ponderadas, de análises estatísticas, da consolidação de valores ou da violação de limites impostos [Smith, 2004a].

## 5.3 Enunciação de propostas

Várias investigações têm sido desenvolvidas no âmbito das métricas que contribuem, em última instância, para a avaliação e análise da qualidade dos dados num SDW. Um conjunto de investigações abarca a definição de métricas para a avaliação da qualidade dos dados fornecidos pela generalidade dos sistemas informáticos e adaptáveis a SDWs [Kahn et al., 2002] [Lee et al., 2001] [Pipino et al., 2002] [Müller & Freytag, 2002] [Parssian et al., 1999]. No domínio dos SDWs, podemos assistir a um grupo de investigações que se debruçam sobre a manutenção e qualidade das actividades decorrentes do processo de ETL e que são corporizadas pelos metadados [Kimball, 2000] [Kimball & Caserta, 2004] [Bouzeghoub & Peralta, 2004] [Amaral, 2003]. Um outro conjunto de investigações tem-se centrado em disponibilizar modelos conceptuais assentes em métricas capazes de aferir sobre a qualidade dos diversos componentes do SDW e a forma como são correspondidas as expectativas dos diversos intervenientes do sistema [Vassiliadis, 2000] [Helfert & Radon, 2000] [Cappiello et al., 2004]. Um último conjunto de estudos ocupa-se da definição de métricas capazes de avaliar a qualidade do modelo multidimensional que serve de suporte ao

SDW [Calero et al., 2001] [Serrano et al., 2002]. Algumas destas propostas apresentam traços de complementaridade entre si e por isso, a sua integração constitui-se num meio imprescindível na gestão coerente dos dados, em especial, na auscultação sobre os dados residentes no SDW.

### 5.3.1 Paradigmas de medição

A generalidade dos modelos de aferição e melhoria da qualidade dos dados recorrem ao paradigma GQM [Basili et al., 1994], como método de elaboração de instrumentos de medida. Posteriormente, um novo método designado por *Measurement Model Life-Cycle* (MMLC) [Cantone & Donzelli, 1998] foi desenvolvido com o intuito de aperfeiçoar o GQM [Piattini et al., 2001]. A importância central destes métodos decorre da sua aplicação nas propostas enunciadas.

#### Goal Question Metric

O GQM [Basili et al., 1994] foi originalmente desenvolvido para a medição da qualidade de *software* e proporciona um modelo para a definição de métricas orientadas aos objectivos, ou seja, possibilita a derivação de medidas a partir de objectivos de medida. O GQM afirma-se como o paradigma no que respeita à construção de métricas conforme é dado a observar num vasto conjunto de investigações [Piattini et al., 2001] [Vassiliadis, 2000] [Bobrowski et al., 1999] [Vassiliadis et al., 1999]. O relevo do GQM resulta da sua simplicidade e por seguir uma aproximação *top-down*. O GQM compreende três níveis: o nível conceptual para definir os objectivos; o nível operacional para definir as questões e o nível quantitativo para derivar as métricas (figura 5-1) [Bobrowski et al., 1999]. O GQM pressupõe duas assumpções. A primeira respeita ao programa de medida não se basear em métricas, mas em objectivos. A segunda aponta a necessidade de adaptação na definição dos objectivos e medidas para cada organização. A utilização da aproximação GQM permite gerar um conjunto de métricas de utilidade claramente justificável [Piattini et al., 2001] [Basili et al., 1994].

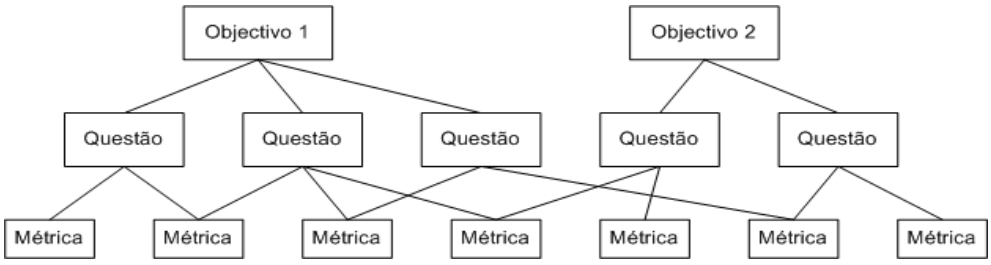



**Figura 5-1 – A aproximação *Goal Question Metric* [Basili et al., 1994].**

Um objectivo é definido para um objecto mensurável: produtos, processos e recursos. Os requisitos de alto nível dos utilizadores podem ser definidos como objectivos. Os objectivos devem ser

definidos em função dum propósito (e.g. avaliar), segundo uma perspectiva (e.g. consumidor) e sobre um objecto (e.g. DM de compras) [Bobrowski et al., 1999]. As questões contribuem para uma maior especificidade dos objectivos e sugerem as métricas relevantes para os objectivos. As questões são usadas para caracterizar a maneira de avaliar um objectivo específico. A resposta às questões permite a avaliação do objecto relativamente ao cumprimento das características a respeitar. O relacionamento entre os objectivos e as métricas é estabelecido através de questões de qualidade. As métricas compõem a sequência das questões em vista aferir quantitativamente o cumprimento das características previstas, ou seja, são valores que expressam algumas medidas das características do objecto. O processo GQM é constituído pelas seguintes fases [Basili et al., 1994] [Vassiliadis, 2000] [Vassiliadis et al., 1999]:

- A identificação de um conjunto de objectivos de qualidade (e.g. a satisfação dos clientes).
- A obtenção das questões a partir dos objectivos e suportadas em modelos de medição de objectos, isto é, devem ser derivadas questões que definam esses objectivos.
- A especificação das medidas que precisam de ser alcançadas de modo a responderem às questões e no sentido da conformidade dos produtos e processos com os objectivos.
- O desenvolvimento de mecanismos de recolha de dados, incluindo de análise e validação.

### Measurement Model Life-Cycle

A aproximação MMLC [Cantone & Donzelli, 1998] baseia-se no GQM e pretende colmatar algumas lacunas entretanto diagnosticadas nesse modelo, nomeadamente, a pouca clareza no processo de geração das métricas e a necessidade de existência de suportes para a criação das questões. É, igualmente, efectuada uma melhor estruturação de cada uma das fases do GQM [Piattini et al., 2001]. A proposta MMLC assenta na definição de métricas, neste estudo designadas por modelos de medida. Os modelos de medida podem ser diferenciados com base na descrição ou na predição. Os modelos de medida descritivos definem um atributo do mundo real (produto, processo ou recurso). A definição é dada em termos de medidas directas ou indirectas sobre uma entidade. Estas últimas incidem sobre outros atributos da entidade ou de outras entidades. Um modelo de medida descritivo pode conter: as propriedades do atributo a medir, um modelo da entidade, uma escala e o mapeamento da entidade entre o mundo empírico e mundo formal. Os modelos de medida preditivos relacionam-se com um atributo duma entidade ainda não disponível, mas que supostamente estará disponível [Cantone & Donzelli, 1998].

As investigações iniciais conduziam o MMLC em torno de três etapas principais: a identificação do modelo, que relaciona o processo de derivação de modelos de medida orientados para as neces-

sidades da organização; a criação do modelo, que consiste na definição e na validação do modelo de medida e a acreditação do modelo, que se ocupa da utilidade prática e da manutenção do modelo de medida [Cantone & Donzelli, 1998].

Posteriormente, em resultado do aprimoramento do MMLC foram estabelecidas quatro etapas necessárias: a identificação, a criação, a aceitação e a acreditação. A etapa de identificação do plano de medida compreende os objectivos e as hipóteses de solução como elementos chave. Os objectivos correspondem aos fins últimos das organizações, podem ser classificados hierarquicamente com base no âmbito (estratégia, área e projecto) e resultam de factores externos e internos. Os objectivos podem ser decompostos em sub-objectivos mais exactos e fáceis de alcançar. O processo que medeia os objectivos e os modelos de medida evolui continuamente em etapas de refinamento claramente identificadas. A cada nova fase de refinamento, um objectivo pode estar associado a modelos de medida. Os objectivos representam a primeira etapa para a identificação de um modelo de medida e são usados para: fornecer as hipóteses, acreditar o modelo e aceitar ou rejeitar o modelo proposto. As hipóteses de solução consistem nos diferentes modos de resolução do problema e são identificadas segundo dois níveis hierárquicos: alto nível ou orientadas ao objectivo e baixo nível ou orientadas aos modelos de medida. As hipóteses orientadas aos objectivos apontam para a estratégia a seguir através da identificação de modelos de medida a adoptar. As hipóteses orientadas ao modelo de medida refinam os requisitos do modelo de medida, incluindo informações específicas (descritivas ou preditivas) do próprio modelo de medida [Cantone & Donzelli, 1998, 1999].

Em seguida, a etapa da criação dos modelos de medida corresponde a um processo estático composto por três elementos sequenciais: a definição das hipóteses orientadas ao modelo de medida, a definição do modelo de medida e a validação do modelo de medida. A primeira componente enriquece e refina os requisitos do modelo de medida, pela especificação de informações sobre a natureza deste (e.g. quando se trata de um modelo composto, os atributos resultam da agregação de outros atributos, gerando novos modelos de medida). A segunda componente consiste na definição dos elementos do modelo de medida. Esta definição depende da natureza do próprio modelo. Para modelos de medida descritivos torna-se necessário: a selecção de um modelo da entidade que mostre as características relevantes da entidade; a definição das propriedades de medida e a escolha de uma função de mapeamento. Enquanto, para modelos de medida preditivos, além das etapas descritas para modelos de medida descritivos, interessa ainda a escolha de um sistema estatístico. Por último, a componente de validação do modelo de medida conclui a etapa da criação do modelo de medida. O processo de validação do modelo de medida depende, igualmente, da natureza do modelo. A validação de um modelo de medida descritivo consiste num

processo formal e na execução duma validação baseada na experimentação. Enquanto, a validação de um modelo de medida preditivo inclui, além do processo formal, também processos estatísticos [Cantone & Donzelli, 1998, 1999].

A etapa referente à aceitação do modelo consiste na execução de uma experimentação sistemática do modelo de medida. Este é aplicado a um contexto simulado das características do ambiente real, com os utilizadores reais de modo a verificar o desempenho obtido em comparação com os objectivos e requisitos iniciais. Trata-se de observar o comportamento do modelo de medida num ambiente de aplicação simplificado e consequentemente, possibilitar uma forma de refinar e enriquecer o conhecimento empírico que o modelo de medida tenta captar [Cantone & Donzelli, 1999].

Por último, a etapa de acreditação do modelo representa o dinamismo processual, isto é, opera com ciclo de vida dos modelos de medida. Esta etapa pretende captar as acções correctivas visando a melhoria contínua da utilidade do modelo em prol da concretização dos objectivos. Assim, esta etapa explica-se pela reorientação dos objectivos inicialmente propostos ou pelo facto do modelo de medida revelar-se um instrumento de pouca utilidade e carecer de remodelação [Cantone & Donzelli, 1998].

### 5.3.2 Propostas gerais de avaliação da qualidade dos dados

As propostas, seguidamente apresentadas, referem-se a modos de aferição da qualidade dos dados aplicáveis à generalidade dos sistemas de informação. Estas propostas pretendem, de maneira geral, averiguar a qualidade dos dados em torno das suas dimensões. A aplicação ao domínio dos SDWs é possível de realizar porque, por um lado, são consideradas as diferentes perspectivas inerentes ao conceito de qualidade, colocando em plano de destaque os intervenientes com o sistema, em especial, os utilizadores finais. Por outro lado, transmitem orientações importantes na condução de iniciativas de melhoria da qualidade dos dados organizacionais.

#### Product and Service Performance for Information Quality

A investigação *Product and Service Performance for Information Quality* (PSP/IQ) [Kahn et al., 2002] surge da necessidade crítica da construção duma metodologia capaz de avaliar o desenvolvimento e entrega dos PIs aos consumidores. A existência duma metodologia que combata esta lacuna permite efectuar aperfeiçoamentos da qualidade dos dados, através de *benchmark* com as outras organizações e como base para diligências visando a melhoria da qualidade dos dados.

As dimensões adoptadas pelo modelo PSP/IQ têm origem na investigação [Wang et al., 1994]. O modelo agrupa as referidas dimensões em torno duma matriz 2 x 2, conforme ilustra a tabela 5-1. Os quatro quadrantes da matriz representam os aspectos relevantes da qualidade dos dados merecedores de melhorias. As colunas da matriz correspondem a duas traves mestras no que diz respeito à qualidade: a conformidade com as especificações e em alcançar ou exceder as expectativas dos consumidores. A conformidade com as especificações assegura que os produtos se encontram ausentes de defeitos. Enquanto que alcançar ou exceder as expectativas dos consumidores alerta para o facto da mera conformidade com as especificações ser insuficiente e por isso, revela-se necessário atingir ou ultrapassar as expectativas e anseios dos consumidores. As linhas da matriz pretendem conferir a natureza das dimensões consideradas: a qualidade do produto e a qualidade do serviço. A qualidade do produto inclui as dimensões tangíveis e relativas às características do produto. Ao passo que a qualidade do serviço corresponde às dimensões intangíveis e relacionadas com os processos de serviço de disponibilização dos dados. O modelo consolida as dimensões em quatro quadrantes:

- Informação idónea: as dimensões deste quadrante são tangíveis e independentes da tarefa e da decisão. As características da informação divulgada devem atingir os standards predefinidos (e.g. os consumidores exigem dados exactos e bem representados).
- Informação fidedigna: as dimensões não podem ser medidas antecipadamente, são apenas avaliadas após ocorrerem e resultam dos processos de conversão dos dados em indicadores de decisão cumprirem os standards previstos. A informação fidedigna deve ser corrente, segura e fornecida atempadamente para servir de suporte à tomada de decisão.
- Informação proveitosa: as dimensões presentes neste quadrante apresentam características de dependência às tarefas e por isso, carecem de contextualização. A informação fornecida deve atingir ou exceder as necessidades dos consumidores.
- Informação usável: estas dimensões distinguem a qualidade dos serviços. Esta apreciação apenas poderá ser executada na óptica dos consumidores. O processo de conversão dos dados em indicadores para os decisores atinge ou excede os anseios dos consumidores. A informação só se torna usável se o consumidor for capaz de a aceder e moldar às suas necessidades.

| | Conformidade com as especificações | Corresponde ou excede as expectativas |
|---|---|---|
| **Qualidade do Produto** | **Informação idónea**:<br>Representação concisa,<br>Representação consistente,<br>Completa e<br>Exacta. | **Informação proveitosa**:<br>Quantidade apropriada,<br>Relevante,<br>Compreensível,<br>Interpretável e<br>Objectiva. |
| **Qualidade do Serviço** | **Informação fidedigna**:<br>Oportuna e<br>Segura. | **Informação usável**:<br>Acessível,<br>Fácil de operar,<br>Reputada e<br>Credível. |

**Tabela 5-1 – O Modelo PSP/IQ [Kahn et al., 2002].**

## AIM Quality

O modelo *AIM Quality* (AIMQ) [Lee et al., 2002] visa a apreciação da qualidade dos dados tendo em vista fornecer uma plataforma rigorosa e pragmática que permite a avaliação e o *benchmark* da qualidade dos dados. A metodologia consiste em três componentes: um modelo de qualidade dos dados, um questionário de medida da qualidade dos dados e técnicas de análise para interpretar as medidas da qualidade dos dados. O primeiro componente, considera o modelo PSP/IQ [Kahn et al., 2002] e procura entender o significado dos dados para os diversos intervenientes.

O segundo componente analisa os componentes da informação e consiste num instrumento de medida de cada uma das dimensões da qualidade dos dados. Este instrumento baseia-se num questionário de medição da qualidade dos dados através da avaliação das dimensões consideradas importantes pelos intervenientes. A construção deste instrumento baseia-se em métodos standard para o desenvolvimento de questionários. A junção das várias dimensões possibilita a avaliação da qualidade dos dados para cada um dos quadrantes do modelo PSP/IQ.

O último componente consiste no recurso a duas técnicas de análise para interpretar as avaliações apresentadas pelo questionário. Estas técnicas usam os quadrantes do modelo PSP/IQ como valores de entrada, analisam e comparam os valores oriundos dos componentes anteriores, tanto o PSP/IQ como o inquérito efectuado, de forma a reconhecer problemas ao nível da qualidade dos dados. As técnicas são iniciadas pela análise a um quadrante do modelo PSP/IQ específico e em seguida são continuadas pelos restantes quadrantes. A primeira técnica consiste no *benchmark* da qualidade dos dados pela comparação com outras organizações e as melhores práticas organizacionais. A segunda técnica efectua o cruzamento e a confrontação entre os resul-

tados alcançados e os intervenientes do sistema. Os resultados obtidos pelo emprego destas técnicas pretendem suportar as decisões sobre as actividades prioritárias para iniciativas de melhoria da qualidade dos dados.

## A proposta de Pipino et al.

A investigação [Pipino et al., 2002] apresenta algumas avaliações objectivas e subjectivas da qualidade dos dados. A proposta expõe quatro formas funcionais (ratio simples, máximo, mínimo e a média ponderada) no auxílio da formulação de métricas objectivas sobre a qualidade dos dados. Por último, é mostrado um modelo que combina as métricas objectivas e subjectivas. A proposta faz uso das dimensões apresentadas em [Wang et al., 1994]. O desenvolvimento de métricas objectivas inicia-se pela definição de uma dimensão da qualidade dos dados. A natureza de cada dimensão individualiza a utilização duma métrica baseada numa função. A primeira, o ratio simples, permite calcular o número de elementos a medir pelo número total de elementos e obtém um valor que se situa entre os limites zero e um. Esta função pode ser aplicada às dimensões: exactidão, completude, consistência, representação concisa, relevância e facilidade de operação.

As funções máximo e mínimo são predominantemente utilizadas em casos que exigem a agregação de vários indicadores ou variáveis sobre a qualidade dos dados. A função mínimo revela-se protectora sobre o valor obtido porque devolve o valor referente ao indicador da qualidade dos dados mais fraco, sendo comum a sua aplicação nas dimensões: credibilidade e quantidade de apropriada dos dados. A utilização da média ponderada sobre alguns indicadores é uma alternativa à utilização da função mínimo. O recurso à média ponderada justifica-se desde que exista um profundo conhecimento sobre os indicadores ou variáveis envolvidas. Por seu turno, a função máximo é usada quando é pretendido um valor que balize o limite máximo sobre um conjunto de indicadores (e.g. a oportunidade e a acessibilidade dos dados).

Após a determinação das avaliações objectivas e subjectivas para cada uma das dimensões realiza-se uma comparação entre os valores obtidos, de modo a detectar discrepâncias ou concordâncias. O modelo recorre a uma matriz 2 x 2 composta por quatro quadrantes, ilustrada na figura 5-2. O objectivo final, para cada uma das dimensões, consiste em posicionar, simultaneamente, os valores das métricas objectivas e subjectivas no quarto quadrante. Sempre que tal não seja conseguido deve-se analisar as razões de tal posicionamento e proceder à tomada de medidas correctivas.

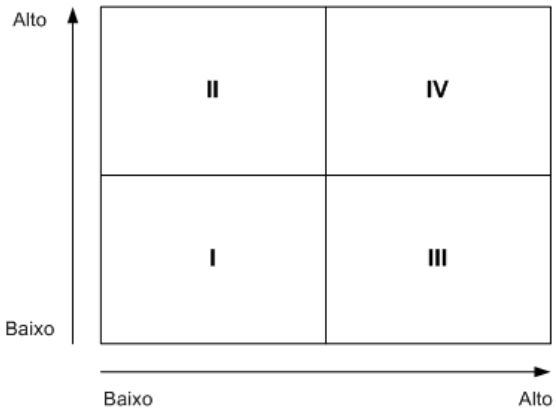



**Figura 5-2 – Comparação entre as métricas objectivas e subjectivas [Pipino et al., 2002].**

### 5.3.3 Propostas orientadas à qualidade dos dados em SDWs

As propostas, a seguir apresentadas, focalizam a sua acção em ambientes de DW e mais especificamente, sobre a qualidade dos dados constantes nesses sistemas. Os modelos de avaliação seguem uma orientação conceptual na aferição dos dados dum SDW. Porém, relevam a importância deste assunto nas actividades comuns dos SDWs.

#### A proposta de Cappiello et al.

Em [Cappiello et al., 2004] é apresentada uma proposta baseada no princípio fundamental que a avaliação sobre as dimensões da qualidade dos dados deve considerar o grau de satisfação das expectativas dos utilizadores. As expectativas dos utilizadores são assunto transversal ao próprio conceito de qualidade. Neste contexto, o estudo propõe um modelo que reúna a fase de avaliação e os requisitos dos utilizadores. A questão central deste modelo consiste em considerar as avaliações dos dados para além dos valores. A utilização específica dos dados relevantes por parte dos utilizadores é assumida como uma componente merecedora de atenção. Por exemplo, um sistema de dados que contenha dados incompletos pode resultar numa avaliação negativa sobre uma métrica objectiva que meça a completude dos dados. Contudo, na óptica dum utilizador ou duma classe de utilizadores específica, o suporte de dados pode conter todos os dados necessários para a execução da tarefa em mãos. O estudo recorre a algumas considerações avançadas em [Pipino et al., 2002], em especial, no que respeita às funções que servem de base para as métricas sobre as dimensões da qualidade dos dados.

A proposta presume que a definição de classes de utilizadores ou intervenientes, possuidores de características próprias, induz um avanço na resolução da problemática em causa, porque considera que as organizações oferecem diferentes tipos de serviços para satisfazer diferentes requisi-

tos dos utilizadores. Assente neste pressuposto, o estudo sugere a efectiva avaliação de uma porção ou subconjunto de dados orientados para a prestação de um serviço específico, em vez de efectuar uma avaliação da qualidade dos dados de todo o sistema em causa. Assim, a investigação apresenta um modelo que recorre a técnicas de atribuição de perfis aos requisitos dos utilizadores. Estas técnicas descrevem estruturadamente os utilizadores e as suas preferências e posteriormente, o modelo usa essa informação na avaliação da qualidade dos dados. Geralmente, cada classe contém utilizadores com características similares, situação que possibilita os utilizadores pertencentes a uma classe específica terem acesso ao mesmo conjunto de serviços. Todavia, é possível que certos utilizadores duma classe possuam requisitos próprios além dos especificados pela classe. Assim, o modelo prevê que cada utilizador estabeleça os seus próprios limites de aceitação em cada uma das dimensões dos dados avaliadas (e.g. se o requisito duma classe para uma dimensão como a oportunidade dos dados é de 0,60 é possível um maior ou menor grau no cumprimento desta exigência por um utilizador particular).

A arquitectura que sustenta o modelo é composta por três módulos principais, conforme ilustra a figura 5-3: o módulo de selecção, o módulo de validação da qualidade e o módulo de perfil dos utilizadores. O utilizador requisita um serviço, e este envia o pedido dos dados ao módulo de selecção, que é responsável pela satisfação desse pedido e pelos metadados necessários ao processo de avaliação dos dados. O módulo de selecção determina os dados possíveis de aceder e a manutenção dos valores das dimensões da qualidade dos dados que têm de ser facultados com a informação requerida. Os dados encontram-se num repositório e o repositório da qualidade contém os metadados sobre a qualidade dos dados. Os dados retornados pelo módulo de selecção são enviados via o módulo de validação da qualidade dos dados, que identifica a função de avaliação dos dados tendo em conta as características contidas no módulo de perfil de utilizador e relativas à classe do utilizador. A função de avaliação é definida pelo estabelecimento da métrica a adoptar para uma dimensão dos dados específica. Os valores da qualidade dos dados resultantes do procedimento de validação são comparados com os níveis previamente estabelecidos para a classe de utilizadores e com as especificações de aceitação particulares a um utilizador.

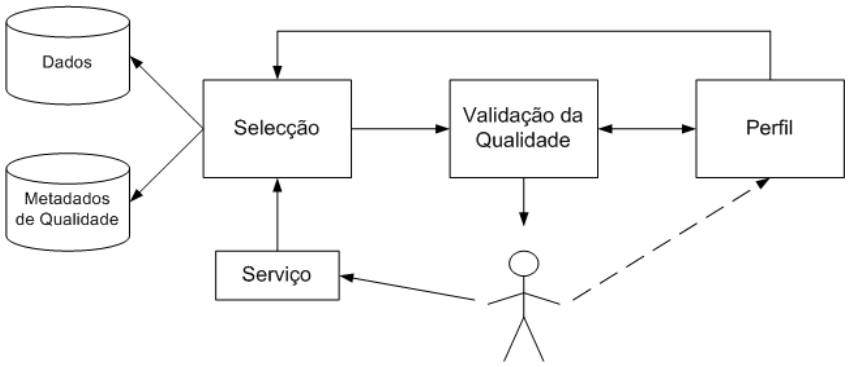



**Figura 5-3 – O modelo da validação da qualidade dos dados pelos utilizadores [Cappiello et al., 2004].**

## A proposta de Bouzeghoub & Peralta

A investigação [Bouzeghoub & Peralta, 2004] apresenta um modelo que analisa e define métricas sobre alguns factores que influenciam o grau de frescura dos dados existentes num sistema informático. A frescura dos dados é reconhecidamente um dos aspectos mais críticos em SDWs. As contínuas exigências de informação atempada e actualizada são assumidas como questões centrais para o sucesso dos SDWs e por isso, são objecto de preocupação pelas organizações e um dos temas em destaque no campo da investigação. A detecção do perfil da frescura dos dados constantes num DW permite avaliar os dados relativamente à idade destes no sistema, a eventuais alterações sintácticas ocorridas e aos ajustamentos nas regras de negócio. A frescura dos dados pode ser percepcionada como um conjunto de factores representativos e possuidores de métricas próprias. Doutro modo, a frescura dos dados pode ser perspectivada como uma dimensão que compreende e agrega factores ou sub-dimensões associadas, como sejam a oportunidade e a actualidade dos dados. O factor actualidade corresponde ao intervalo de tempo entre a extracção dos dados das fontes e a sua disponibilização aos utilizadores. Enquanto, o factor oportunidade é entendido como quantas vezes os dados mudam ou quantas vezes novos dados são criados nas fontes (tabela 5-2).

| Factor | Métrica | Definição |
|---|---|---|
| Actualidade | Actualidade | O tempo que decorre desde que os dados saíram da fonte (a diferença entre o tempo de consulta e o tempo de extracção). |
| | Obsolescência | O número de actualizações na fonte desde o tempo de extracção. |
| | Taxa de frescura | A percentagem das linhas que estão actualizadas na vista. |
| Oportunidade | Oportunidade | O tempo passado desde a última actualização na fonte (a diferença entre o tempo de consulta e o tempo da última actualização). |

**Tabela 5-2 – Factores e métricas sobre a frescura dos dados [Bouzeghoub & Peralta, 2004].**

A investigação alerta para o facto da frescura dos dados enviados aos utilizadores depender, obviamente, da frescura dos dados extraídos, mas igualmente dos processos de extracção, integração e entrega dos dados. A investigação considera os SDWs como sistemas que materializam vistas sobre dados refrescados periodicamente. A implementação seguida para a materialização das vistas sobre os dados, entre os diversos processos de fluxo de dados no SDW, poderá influenciar decisivamente a frescura dos dados entregues aos utilizadores. O processo de refrescamento dos dados nos SDWs pode respeitar políticas de *pull* ou *push* e assumir uma postura síncrona ou assíncrona (assunto anteriormente referido). Ora, dado que os dados nos SDWs provêm de fontes heterogéneas e dispersas entre si, é comum optar pela conjugação de políticas complementares e a introdução de assíncronismos entre os processos, situação que induz a criação de tempos de espera. Assim, a materialização de vistas aumenta potencialmente as inconsistências entre as fontes e o DW e por isso, o problema na manutenção das vistas consiste na actualização das vistas em resposta às mudanças ocorridas nas fontes.

Além do processo de refrescamento, a natureza dos dados constantes nas vistas materializadas influencia, directamente, a aferição dos dados. Os dados podem apresentar características de estabilidade ou de mutabilidade (frequentes ou raras). Em SDWs, na manutenção de dados frequentemente modificáveis, revela-se conveniente a minimização da entrega de dados desactualizados, bem como a medição do tempo em que os dados não sofrem mudanças. Enquanto, na manutenção de dados estáveis ou raramente mutáveis interessa questionar as vezes que novos dados são criados e a idade destes no sistema. Nesse sentido, a avaliação dos dados tendo em conta a sua natureza condiciona a aplicação de métricas adequadas a cada caso (tabela 5-3).

| Mudanças frequentes | Mudanças a longo prazo | Estáveis |
|---|---|---|
| Actualidade<br>Obsolescência | Oportunidade | Oportunidade |

**Tabela 5-3 – Relação das métricas e tipos de dados em SDWs [Bouzeghoub & Peralta, 2004].**

## A proposta de Vassiliadis

A investigação [Vassiliadis, 2000] apresentada no capítulo anterior, no âmbito da melhoria da qualidade dos dados em ambientes de DW, prossegue com a especialização do meta-modelo, através do enriquecimento por padrões e modelos para objectivos de qualidade e métricas para casos concretos da gestão da qualidade. Neste campo particular, interessa revelar a extensão do meta-modelo da qualidade pela aplicação de modelos de métricas. Assim, as métricas consideradas associam-se aos quatro principais processos operacionais dos DWs: administração e desenho; implementação e avaliação de *software*; carregamento dos dados e utilização dos dados. Adicio-

nalmente, é integrado no modelo de qualidade a qualidade dos dados constantes no sistema. Iremos observar apenas as métricas relativas à qualidade e utilização dos dados por serem aquelas mais directamente centradas na problemática focada no nosso estudo.

A natureza inerente aos SDWs faz transparecer a necessidade deste sistema potenciar um conjunto de funcionalidades úteis aos consumidores finais, em especial, as relativas à acessibilidade e utilidade dos dados. A dimensão acessibilidade pretende agrupar um conjunto de factores de qualidade relacionados com o acesso aos dados: a segurança, a disponibilidade do sistema e a disponibilidade transaccional. A dimensão utilidade descreve as características temporais dos dados, bem como, os níveis de resposta do sistema: a capacidade de resposta do sistema, a oportunidade dos dados e a interpretação dos dados (tabela 5-4).

| Dimensão | Métrica | Definição |
|---|---|---|
| Acessibilidade | Disponibilidade do sistema | A percentagem de tempo em que os dados não se encontram disponíveis devido a falhas do sistema. |
| | Disponibilidade transaccional | A percentagem de tempo em que os dados não se encontram disponíveis devido a operações de actualização. |
| | Segurança | Nº de procedimentos de autorização não documentados. |
| Utilidade | Resposta do sistema | Nº de processos não listados aos utilizadores. |
| | Oportunidade: actualidade | Nº de dados necessários não presentes em tempo de transacção. |
| | Oportunidade: volatilidade | Nº de dados necessários não presentes em tempo válido. |
| | Interpretação | Nº de máquinas ou software não documentados;<br>Nº de dados não documentados; |

**Tabela 5-4 – Factores e métricas sobre a qualidade de uso (adaptado: [Vassiliadis, 2000]).**

No que respeita à qualidade dos dados armazenados nos DWs, a investigação mostra que esta é influenciada pelos diversos processos inerentes ao próprio sistema. A proposta adopta a definição do conceito de qualidade dos dados divulgada em outras investigações [Strong et al., 1997] [Wang et al., 1994], incluindo a: completude, credibilidade, exactidão, consistência e interpretação dos dados (tabela 5-5).

| Dimensão/Métrica | Definição |
|---|---|
| Completude | A percentagem detectada de dados armazenados incompletos relativamente ao mundo real. |
| Credibilidade | Percentagem de dados incorrectos fornecidos por cada fonte. |
| Exactidão | A percentagem detectada de dados armazenados incorrectos relativamente aos valores do mundo real, devido a problemas de captura dos dados. |
| Consistência | A percentagem detectada de dados armazenados inconsistentes. |
| Interpretação | Número de dados não totalmente descritos. |

**Tabela 5-5 – Factores e métricas sobre a qualidade dos dados (adaptado: [Vassiliadis, 2000]).**

## Data quality management

A proposta *Data Quality Management* (DQM) [Helfert, 2001] consiste na construção de um método baseado na gestão da qualidade dos dados e que se enquadra nas investigações desenvolvidas pelo *Competence Center 'Data Warehousing Strategy (CCDWS)*[6]. O método toma por base a definição de cinco dimensões inerentes ao conceito de informação: a semiótica (pragmática, semântica e sintáctica), o meio, a idoneidade, a oportunidade e a novidade. A informação é vista como um subconjunto do conhecimento. É igualmente considerada a visão clássica do conceito qualidade. Assim, a qualidade é definida como a adequação ao uso e assente em dois estabilizadores principais. Primeiro, a qualidade de desenho que consiste em dotar o produto das características apropriadas em vista satisfazer os desejos dos consumidores. Portanto, corresponde aos requisitos de informação e ao desenho do PI. Segundo, a qualidade de conformidade que resulta da elaboração do produto conforme o especificado e ausente de defeitos.

Esta proposta assenta na TQM, como plataforma metodológica estável e aplicável à gestão dos dados. A metodologia compreende uma política da qualidade, um planeamento da qualidade, um controlo da qualidade, uma segurança da qualidade e um melhoramento da qualidade. A intenção da presente proposta passa pela definição duma estrutura integrada de todas as actividades organizacionais, pela atribuição de papéis e responsabilidades bem definidas, no sentido de promover a melhoria contínua do PI. O ciclo do processo da TQM (PDCA) constitui-se como o ponto fundamental para a melhoria contínua da qualidade na organização.

Paralelamente, o estudo promove a reorganização de um conjunto de dimensões de qualidade e de métricas de medida, contextualizando-as a SDWs. Assim, é disposto um conjunto relevante de características de qualidade em três níveis de semiótica (pragmático, semântico e sintáctico) e nos dois componentes de qualidade (qualidade de desenho e qualidade de conformidade). O nível pragmático é responsável pelo entendimento de todas as características relevantes nos processos dos dados e nas informações divulgadas aos utilizadores. Neste nível, na qualidade de desenho predominam as dimensões relevância e completude dos dados. Enquanto, a qualidade de conformidade salienta a oportunidade, actualidade e eficiência dos dados. O nível semântico responde perante o significado dos dados. A qualidade de desenho caracteriza-se pela definição dos dados de modo preciso, objectivo e compreensível. A qualidade de conformidade reconhece a importância dos factores: interpretação, exactidão, credibilidade, consistência e completude dos dados. O nível sintáctico compreende as preocupações sintácticas dos caracteres e símbolos usados. A

[6] Fundado pela Universidade de *St. Gallen*, Suiça, em Janeiro de 1999. O CCDWS deu origem ao *Competence Center 'Data Warehousing 2'*.

qualidade de desenho ocupa-se da consistência e da sintaxe adequada. Ao passo que a qualidade de conformidade abrange a exactidão sintáctica, a representação consistente, a segurança e a acessibilidade dos dados (tabela 5-6).

| | **Pragmático** | **Semântico** | **Sintáctico** |
|---|---|---|---|
| **Qualidade de desenho** | Relevante<br>Completa | Precisa<br>Objectiva<br>Compreensível | Consistente<br>Sintaxe adequada |
| **Qualidade de conformidade** | Oportuna<br>Actual<br>Eficiente | Interpretável<br>Exacta<br>Credível<br>Consistente<br>Completa | Exactidão sintáctica<br>Representação consistente<br>Segura<br>Acessível |

**Tabela 5-6 – Características dos dados de acordo o nível semiótico e a qualidade.**

O modelo propõe, igualmente, formas de medida enquadradas segundo o nível de semiótica considerado. Relativamente ao nível pragmático, as medidas devem sustentar-se em informações sobre os processos e aplicações do sistema. A nível semântico deve recorrer-se à comparação com o mundo real e à experiência acumulada (histórico). Por último, a nível sintáctico deve proceder-se à comparação com standards sintácticos e acordos predefinidos.

Em suma, encontram-se reunidas as condições necessárias para o planeamento e medida da qualidade dos dados em ambientes de DW. O modelo de qualidade DQM compõe-se em três patamares, conforme mostra a figura 5-4. O patamar superior expõe os requisitos de qualidade subjectivos sobre os processos de informação e a informação aos utilizadores. O patamar intermédio compreende um modelo integrado que representa os requisitos específicos de qualidade num nível lógico. Este patamar compreende a especificação das características da qualidade dos dados e de medidas objectivas de avaliação. Os requisitos estruturados dos utilizadores correspondem a objectivos de qualidade a alcançar. Por fim, o patamar inferior representa as técnicas de avaliação e os sistemas de medida para validar as qualidades especificadas. Geralmente, são usadas métricas objectivas e subjectivas. A comparação no tempo dos valores verificados permite induzir comportamentos de tendência e promover iniciativas de melhoria [Helfert & Radon, 2000].

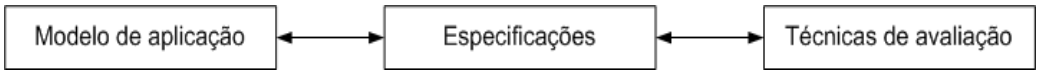



**Figura 5-4 – Modelo de qualidade dos dados [Helfert, 2001].**

### 5.3.4 Propostas de avaliação da qualidade do modelo multidimensional

O modelo multidimensional é consensualmente reconhecido como a plataforma adequada na conceptualização dos dados existentes num DW [Kimball et al., 1998]. A justificação por esta opção reside na facilidade de concretização de análises complexas aos dados e no modo intuitivo de visualização dos mesmos (e.g. a organização das datas das vendas por mês, trimestre e anual) [Chaudhuri & Dayal, 1997] [Serrano et al., 2003]. Em [Kimball et al., 1998] refere-se os bons tempos de resposta nas consultas aos dados e a facilidade de entendimento dos dados e dos metadados pelos intervenientes. Ora, diferentes esquemas dimensionais podem ser obtidos com resultados semelhantes, por isso, importa optar pelo que oferece melhores garantias em termos de eficácia, eficiência e compreensão das estruturas envolvidas. A enunciação de um conjunto de métricas capazes de avaliar estas características pode mostrar-se um meio contributivo para a melhoria da qualidade dos dados disponibilizados aos diversos intervenientes.

#### A proposta de Calero et al.

A investigação [Calero et al., 2001] centraliza a sua acção na construção de métricas objectivas para avaliar o modelo multidimensional em estrela que suporta os dados num DW. A linha condutora subjacente a este estudo passa por assumir a existência de um modelo multidimensional de elevada qualidade, como forma de garantir uma boa qualidade da informação fornecida aos consumidores. Tendo em vista auxiliar os desenhadores dos modelos multidimensionais são construídas métricas capazes de permitir a escolha entre modelos multidimensionais equivalentes. O estudo expõe a importância que os SDWs assumem no contexto organizacional actual e realça a informação de qualidade proporcionada pelos SDWs como um factor crítico de sucesso para a excelência na tomada de decisões estratégicas. Assim, é apresentada uma estrutura conceptual orientada para o modo como se obtém informação de qualidade e que revela duas componentes essenciais: a qualidade do DW e a qualidade da apresentação (figura 5-5).

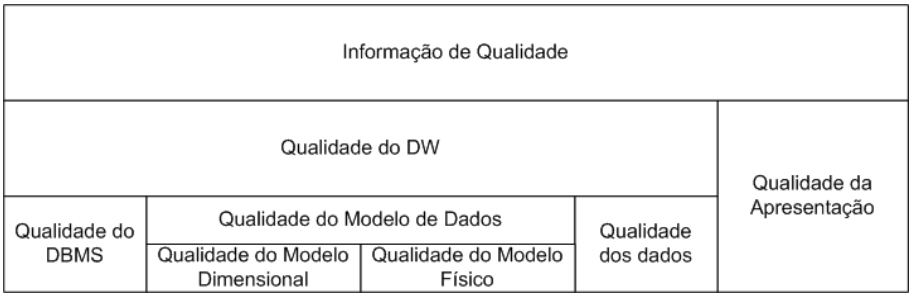



**Figura 5-5 – Componentes da informação de qualidade [Calero et al., 2001].**

A definição das métricas pressupõe o cumprimento de um formalismo metodológico, que assenta num conjunto de etapas capazes de assegurarem a credibilidade das métricas propostas, conforme é observável na figura 5-6. Assim, na definição de uma métrica recorre-se a um método que envolve três actividades principais: a definição, a validação teórica e a validação empírica. A primeira actividade respeita à proposta da métrica e toma em linha de conta o sistema que pretendemos medir e a experiência dos desenhadores do sistema. O recurso à aproximação GQM revela-se útil nesta actividade. A segunda actividade promove uma validação teórica da métrica e consiste na sua validação formal. A última actividade respeita à aprovação prática da utilidade da métrica, que pode ser baseada na experimentação ou em casos de estudo.

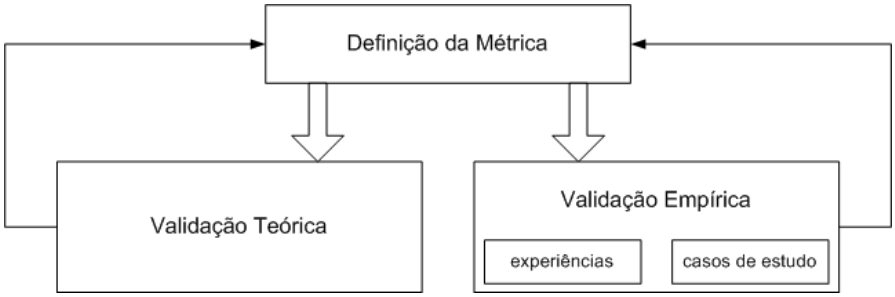



**Figura 5-6 – Etapas para a definição e validação de métricas [Calero et al., 2001].**

As métricas apresentadas no domínio do DW consistem em métricas ao nível da tabela, ao nível da estrela e ao nível do esquema. Quanto ao nível da tabela podem existir métricas que avaliem o número de colunas de uma tabela e número de chaves estrangeiras de uma tabela. Ao nível da estrela, podem ser definidas, entre outras, métricas que determinem o número de dimensões da estrela, o número de tabelas da estrela, o número de colunas das dimensões da estrela, o número de colunas da estrela e o número de chaves estrangeiras da tabela de factos. Quanto ao esquema, é possível construir, por exemplo, métricas que indiquem o número de tabelas de factos do esquema, o número de dimensões do esquema, o número de dimensões partilhadas do esquema e o número de atributos das tabelas de facto do esquema.

## A proposta de Serrano et al.

No seguimento da investigação desenvolvida em [Calero et al., 2001], a proposta [Serrano et al., 2002] surge com o objectivo de efectuar uma validação empírica sobre as métricas, como suplemento à validação formal realizada na investigação anterior. Neste sentido, este estudo centra-se em duas métricas anteriormente definidas e validadas formalmente: o número de tabelas de facto no esquema e o número de dimensões no esquema. A proposta pretende, através da validação empírica (experiências), averiguar se estas duas métricas se constituem como indicadores de

avaliação sobre a dimensão relativa à compreensão dos dados. Assim, averigua a correlação entre as duas métricas e a compreensão do modelo multidimensional dos dados. Estas métricas podem ser usadas por uma classe específica de intervenientes (e.g. os desenhadores do modelo multidimensional) porque fornece indicadores decisivos na opção entre modelos multidimensionais semanticamente equivalentes. O estudo realça que o número de tabelas de facto pode constituir-se como um forte indicador sobre a complexidade do modelo multidimensional, mas que o mesmo não se verifica com o número de tabelas dimensão [Serrano et al., 2002]. Mais precisamente, o estudo considera que a complexidade do modelo dimensional pode ser directamente influenciada pelo número de tabelas de facto, o número total de tabelas e o número de chaves estrangeiras [Serrano et al., 2003].

O método seguido para a definição e validação das métricas foi incorporado pela introdução de estudos na componente da validação empírica e com a adição duma perspectiva psicológica. A justificação para a introdução da componente psicológica consiste em explicar o entendimento da influência dos valores das métricas na compreensão do modelo multidimensional dos dados (figura 5-7).

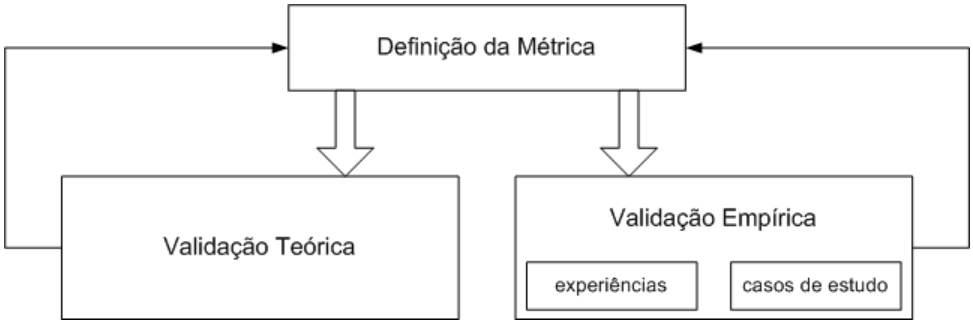



**Figura 5-7 – Etapas da definição e validação de métricas [Serrano et al., 2002].**

## Comparação das propostas

As abordagens apresentadas visam a aferição da qualidade dos dados em SDWs. As propostas apresentam algumas diferenças que podem ser aproveitadas para as complementar entre si. A tabela 5-7 refere-se à comparação entre as propostas apresentadas, em especial, no que concerne à natureza das métricas, ao objecto do DW aplicável e ao nível de concretização.

| Proposta | Natureza | Objecto | Descrição | Nível | Origem |
|---|---|---|---|---|---|
| GQM [Basili et al., 1994] | Métricas | Qualidade do SDWs | Paradigma das métricas da qualidade do software que proporciona a definição de métricas orientadas aos objectivos. Consiste numa aproximação *top-down* assente em 3 níveis: conceptual (objectivos), operacional (questões) e quantitativo (métricas). | Conceptual | |
| MMLC [Cantone & Donzelli, 1998] | Métricas | Qualidade do SDWs | Focaliza as actividades de gestão para a geração, melhoramento e alcance dos objectivos organizacionais. Consiste numa extensão à GQM, apresentando-se mais estruturada em cada fase. Identifica as actividades principais com os *inputs* e *outputs* associados. | Conceptual | GQM [Basili et al., 1994] |
| [Kahn et al., 2002] | Métricas subjectivas | Qualidade dos dados | Modelo PSP/IQ que agrupa as dimensões por: qualidade do produto ou serviço e qualidade de conformidade ou projecto. | Técnico | Dimensões de [Wang et al., 1994] |
| [Lee et al., 2002] | Métricas subjectivas | Qualidade dos dados | Modelo AIMQ que propõe a avaliação e *benchmark* dos dados. O modelo é composto por: PSP/IQ, questionário e técnicas de análise. | Técnico | PSP/IQ [Kahn et al., 2002] |
| [Pipino et al., 2002] | Métricas objectivas e subjectivas | Qualidade dos dados | Modelo que combina as métricas objectivas e subjectivas para detecção de concordâncias e discordâncias. O modelo pressupõe que a natureza de cada dimensão infere o uso de uma métrica. | Técnico | PSP/IQ [Kahn et al., 2002] |
| [Cappiello et al., 2004] | Métricas objectivas | Qualidade de projecto dos dados e metadados | Modelo que reúne a fase de avaliação e os requisitos dos utilizadores. O modelo assenta numa arquitectura composta pelos módulos: selecção, validação da qualidade e *profiling*. | Conceptual | Métricas objectivas [Pipino et al., 2002] |
| [Bouzeghoub & Peralta, 2004] | Métricas objectivas | Frescura dos dados em SDWs | Modelo que analisa e define métricas sobre alguns factores que influenciam o grau de frescura dos dados existentes num sistema: actualidade e oportunidade dos dados. | Conceptual | Dimensões de [Wang et al., 1994] |
| [Helfert, 2001] | Método de gestão da qualidade dos dados | Qualidade do SDW | Método de gestão da qualidade dos dados (DQM) composto por um modelo da qualidade dos dados: os requisitos subjectivos, os requisitos específicos, as medidas de avaliação e os sistemas de medida. Paralelamente, as dimensões são reorganizadas por: qualidade de projecto ou conformidade, as características semióticas e as métricas associadas. | Conceptual | TQM<br>Dimensões de [Wang et al., 1994]<br>QFD |
| [Calero et al., 2001] | Métricas objectivas | Modelo multidimensional do DW | Modelo para a definição e validação (teórica e empírica) de métricas que permitem optar por tabelas, estrelas e esquemas semanticamente equivalentes. | Técnico | GQM [Basili et al., 1994] |
| [Serrano et al., 2002] | Métricas objectivas | Modelo multidimensional do DW | Validação empírica de métricas objectivas (nº de tabelas de facto e de dimensões do esquema), pela observação da correlação entre as duas métricas e a compreensão do modelo multidimensional dos dados. | Técnico | Métricas objectivas [Calero et al., 2001] |

**Tabela 5-7 – Propostas de métricas visando a avaliação da qualidade dos dados.**

## 5.4 Desenvolvimento de um programa de métricas

Conforme referido anteriormente, as métricas devem ser encaradas como indicadores, preferencialmente, objectivos sobre os processos e fluxos de dados decorrentes num SDW. A aplicação das métricas pode manifestar-se em todo o universo dum DW, tendo em vista a prestação de informações precisas e concisas sobre o sistema. Porém, iremos reportarmo-nos, preferencialmente, àquelas que incidem sobre a qualidade dos dados. O principal desígnio consiste em assumir a análise das métricas como uma valiosa ajuda na eventual detecção de potenciais disfunções e não como meio de diagnóstico ou tratamento rigoroso e definitivo. Desde que estrategicamente colocadas em pontos-chave do SDWs e enquadradas num eficaz sistema de gestão da qualidade, elas podem ser usadas no sentido de monitorizar o progresso alcançado, permitindo a identificação de potenciais problemas e desencadear uma acção correctiva em tempo útil [Smith, 2004a, 2004b]. Assim, em vista o cumprimento dos objectivos propostos para as métricas, mostramos, numa primeira fase, os aspectos de gestão merecedores de especial atenção no que respeita à gestão e manutenção das métricas e numa fase posterior a descrição de algumas métricas, comummente aceites, como pertencentes a um lote representativo da qualidade dos dados.

### 5.4.1 Administração de métricas

O desenvolvimento de um programa de métricas deve prever o cumprimento de alguns pressupostos, como sejam o patrocínio por um gestor de topo da organização; a identificação dos objectivos; a confrontação dos custos e benefícios potenciais; a definição dos tipos e níveis; o seguimento de uma metodologia de suporte à formulação de métricas capazes de responder perante as diversas dimensões da qualidade dos dados consideradas e o enquadramento adequado das métricas numa arquitectura de promoção da qualidade. A importância das métricas na aferição dum SDW e em consequência, o contributo implícito que prestam para o sucesso da organização, pressupõe a existência de um patrocínio ao nível mais elevado da estrutura hierárquica. Na equipa de elaboração das métricas deverá constar, igualmente, o administrador do DW, o administrador dos dados e outros supervisores dos dados [Graville, 2004].

Em seguida, deve ser documentada a estrutura e o desenvolvimento das métricas para descrever e assegurar um balanceamento entre os tipos e níveis das métricas. Esta preciosa informação constitui uma componente fundamental dos metadados sobre as métricas definidas. A identificação do objecto a avaliar e a descrição do processo de pormenorização de avaliação deve ser necessariamente mantida. Consideremos, por exemplo, o caso da avaliação do número de erros nas disciplinas frequentadas por 10 alunos, em que cada aluno frequenta 10 disciplinas e que um de-

les possui um erro na atribuição duma disciplina. O resultado obtido indica uma exactidão de 99% nas disciplinas frequentadas, mas apenas 90% nos alunos. Este exemplo salienta a necessidade em determinar qual o nível de granularidade que a métrica deve respeitar, ou seja, neste caso a métrica deverá avaliar os valores referentes às disciplinas frequentadas ou aos alunos?

Ainda sobre a definição do processo de avaliação do objecto alvo, importa salientar que raramente uma única métrica descreve de forma completa e verdadeira a situação (se compararmos a um índice bolsista, podemos observar que o valor do índice serve como mera referência). Muitas vezes só se conseguem tirar conclusões quando várias métricas são analisadas conjuntamente. Considere-se, por exemplo, dois pontos de operação com o sistema informático X e Y com complexidades e dimensões semelhantes. Enquanto no ponto X assiste-se a uma produtividade (inserção) diária na ordem dos 100 linhas por pessoa, no ponto Y são inseridos cerca de 500 linhas por pessoa. As conclusões poderiam ser precipitadas se não se considerasse que no ponto X a taxa de falhas (valores, incorrectos, ausentes ou excesso de valores por defeito) foi de 1 por cada 1000 linhas introduzidas, enquanto esse valor para o ponto Y ascendeu a 15 falhas. Logo, a complementaridade entre métricas e tipos de métricas a subscrever deve ser uma realidade porque permite a confrontação de resultados e consequentemente, um maior rigor na análise das imperfeições verificadas e orientações sobre as causas dessas imperfeições.

Outro ponto alvo de ponderação respeita à análise de custos e benefícios da introdução das métricas. Os custos não respeitam somente à efectivação inicial das métricas, pois permanecem durante todo o ciclo de vida destas. O custo total duma métrica corresponde à soma, entre outros, da contínua recolha, compilação, armazenamento, análise e gestão. No que respeita aos benefícios, estes podem manifestar-se quer nas poupanças de tempo e custos na manutenção dos dados, quer na melhoria da qualidade dos serviços prestados e nas informações divulgadas aos consumidores. Regra geral, os benefícios são de mais difícil determinação do que os custos, por isso, a fixação antecipada de patamares a atingir ou a avaliação por *benchmark* com standards, revela-se de primordial importância na análise do ROI obtido. A opção pelo *benchmark* de standards apresenta, igualmente, um outro efeito porque encoraja a organização a atingir valores de mercado e a não se remeter somente à melhoria dos valores históricos registados (e.g. uma organização define um objectivo que atinge a mediocridade, enquanto a concorrência define um objectivo idêntico para a excelência). É importante compreender o passado e o presente, mas mais importante é entender os níveis capazes de atingir no futuro [Smith, 2004b] [Graville, 2004].

Ainda, em relação à necessidade de cumprimento dos objectivos a atingir pelas métricas é importância balizar convenientemente os limites de valores a partir dos quais são accionados os meca-

nismos de alerta e as rotinas de averiguação sobre os dados com perturbações de qualidade [Wood, 2002]. A definição dos limites a impor deriva, além dos aspectos focados anteriormente, dos requisitos em termos de qualidade dos dados exigidos pelos consumidores dos dados (e.g. a tolerância máxima de desvio sobre as vendas efectuadas diariamente é de 2% ou o tempo de resposta às consultas ser inferior a 3 minutos) [Helfert & Maur, 2001]. A excessiva condescendência relativamente aos valores registados, pode conduzir a falhas nos dados e a eventuais consequências em termos da tomada de más decisões. Também, o estabelecimento de limites muito estreitos pode contribuir para a inoperacionalidade ou estrangulamentos do sistema e o consequente descrédito deste, na medida que inviabiliza a tomada de decisões em tempo útil ou exige a constante necessidade de confrontação de resultados. Logo, a necessária ponderação sobre a abrangência do domínio de valores aceitáveis, bem como, a sintonia entre os objectivos pretendidos pelos diversos intervenientes no DW é um aspecto de importância crucial para o cumprimento pleno de um programa de métricas sobre a qualidade dos dados [Vassiliadis, 2000].

### 5.4.2 Enunciação de métricas

Iremos descrever um conjunto de métricas representativas dos aspectos, vulgarmente, mais característicos e elucidativos da qualidade dos dados. Deste modo, pretendemos dispor de indicadores objectivos sobre os vectores orientadores da qualidade dos dados. As métricas propostas procuram, sobretudo, salientar a perspectiva dos utilizadores relativamente à qualidade dos dados acedidos e que se encontram armazenados nos SDWs. A descrição estrutural das métricas estabeleceu-se segundo o paradigma GQM e assenta em investigações demonstrativas da aplicação do referido paradigma à qualidade dos dados [Bobrowski et al., 1999] e outras que o relacionam ao âmbito dos DWs [Vassiliadis, 2000] [Jarke & Vassiliou, 1997] [Amaral, 2003]. É possível estabelecer um conjunto de dimensões de âmbito transversal sobre o conceito de qualidade dos dados e que decorre das diferentes investigações, mesmo apesar de algumas discordâncias, quer em termos das dimensões consideradas, quer em termos dos significados dessas dimensões [Jarke & Vassiliou, 1997] [Bobrowski et al., 1999] [Wang et al., 1994] [Helfert & Maur, 2001] [Helfert & Radon, 2000] [Strong et al., 1997]. A questão em definir um conjunto de dimensões equilibrado em quantidade e representativo em significado é um assunto focado em diversas investigações [Lee et al., 2002] [Scannapieco & Catarci, 2002] [Lee & Strong, 2004] [Cappiello et al., 2004] [Naumann & Rolker, 2000]. A dificuldade de consenso entre as terminologias adoptadas, noção e alcance dessas dimensões conduz ao necessário acompanhamento dos respectivos significados.

A identificação material das métricas deve resultar, preferencialmente, da situação concreta de cada caso. Assim, deve identificar-se as dimensões a considerar e os critérios que as representam

numa realidade organizacional real. Porém, dada a predominância de algumas dimensões da qualidade dos dados, indiferentemente do contexto de aplicação, fica viabilizada, a ambientes de DW, a derivação de um conjunto nuclear representativo de medidas de avaliação sobre as principais dimensões da qualidade dos dados. O lote representativo das dimensões a considerar, neste relatório, inclui a oportunidade, a completude, a exactidão, a acessibilidade, a relevância e a interpretação. Cada dimensão tem associada uma ou mais métricas que representam os critérios requeridos em termos de qualidade dos dados. Assim, a aplicação das métricas visa indicar o grau de presença dos critérios nos dados. Alguns critérios sobre os dados corporizam uma natureza subjectiva no momento da avaliação e recolha de resultados das métricas, como seja o caso da interpretação dos dados por parte dos consumidores [Cappiello et al., 2004]. Nestes casos, optou-se, além de manter a subjectividade da aferição, em contornar a própria subjectividade através da avaliação objectiva dos aspectos que a podem amenizar (e.g. a existência ou não de documentação e a ajuda sobre os dados divulgados derivados das consultas).

Em seguida, iremos apresentar um conjunto de métricas baseadas no modelo GQM. Em primeiro lugar, procedeu-se à identificação dos objectivos para cada dimensão considerada. Dado que as métricas a definir não respondem perante nenhuma situação concreta, optou-se por considerar como objectivo a atingir, nas diversas métricas, a simples aferição do desempenho das dimensões consideradas. Depois, são definidas as questões sobre os objectivos considerados e por fim, são derivadas as métricas com intuito de fornecer respostas às questões. Adicionalmente, são apresentadas possíveis técnicas de captação dos valores das métricas.

## Exactidão

A exactidão corresponde ao armazenamento correcto dos factos ou valores do mundo real, isto é, consiste em possuir os valores dos dados certos e de confiança [Pipino et al., 2002]. A exactidão pode ser analisada segundo três vectores: o nível sintáctico, o nível semântico e o nível de conteúdo. O primeiro aspecto refere-se ao tipo e domínio dos dados. O segundo aspecto ocupa-se de questões relativas à integridade referencial e às regras de negócio. Por fim, o nível de conteúdo consiste no armazenamento efectivo do valor real. O quociente entre os valores correctos duma fonte e a totalidade de valores na fonte descreve, numa visão abstracta, um modelo de medida para avaliar a exactidão [Naumann & Rolker, 2000]. O propósito da aplicação de métricas sobre a exactidão consiste na avaliação dos dados existentes no repositório (tabela 5-8).

| Dimensão | Questão | Métrica |
|---|---|---|
| Exactidão sintáctica | Os dados respeitam o domínio dos dados? | Percentagem das linhas ou colunas que pertencem ao domínio dos dados |
| Exactidão semântica | As chaves estrangeiras existem nas dimensões? | Percentagem das linhas que contêm chaves estrangeiras nas dimensões |
| Exactidão de conteúdo | Os dados quando comparados com o mundo real estão correctos? | Percentagem das linhas correctas quando comparados com o valor real |

**Tabela 5-8 – Métricas a definir para a dimensão exactidão.**

A técnica de captação para a exactidão sintáctica pode passar pela utilização duma ferramenta de análise dos dados capaz de realçar os valores fora do intervalo de valores previsto. Esta mesma técnica pode ser aplicada igualmente como auxiliar na detecção de valores não correspondentes aos efectivamente existentes no mundo real. A recolha de resultados relativos à integridade referencial pode ser conseguida com a colaboração duma ferramenta de auditoria de dados.

## A frescura dos dados

A frescura dos dados em SDWs assume importância capital na qualidade dos dados divulgados, uma vez que a tomada de decisões encontra-se normalmente condicionada pela questão temporal. Este período de tempo pode ser mais ou menos estreito e caracterizado por abarcar a confluência de diversos factores, muitas vezes, antagónicos. Contextualizando ao tema desta dissertação, a disponibilização de informações no exacto momento que são necessárias pode influenciar o sentido duma decisão e o grau de confiança da sua acção. As informações variam, naturalmente, com o tempo e por isso, mostra-se necessária a actualização do DW, em vista manter a consistência entre os dados armazenados no SO e os mantidos no repositório do DW. A manutenção de um calendário das actualizações, especificando a periodicidade dos carregamentos dos dados num DW deve ser um aspecto a considerar [Amaral, 2003].

O assunto da frescura dos dados, no campo dos SDWs, pode ser avaliado segundo duas vertentes: a actualidade e a oportunidade [Bouzeghoub & Peralta, 2004]. A primeira mede o intervalo de tempo entre a mudança dos dados na fonte sem que essa mudança se reflicta na vista materializada, na prática, em SDW, corresponde à estimativa da diferença entre o tempo de extracção dos dados e o tempo de entrega destes. A actualidade pode ainda ser avaliada segundo a obsolescência, que mede o número de actualizações duma fonte desde o tempo de extracção dos dados e assim estimar o número de frequências de actualização. A outra vertente da frescura dos dados, a oportunidade, é descrita como a medida sobre a extensão da idade dos dados que é considerada apropriada para a tarefa em mãos. É comummente estimada como o tempo passado desde a última actualização da fonte e limitada pela frequência de actualização da fonte (tabela 5-9).

| Dimensão | Questão | Métrica |
|---|---|---|
| Actualidade | Os dados são usados a tempo da tomada de decisão? | Percentagem de decisões tomadas usando os dados armazenados |
| | Quantas vezes, os mesmos dados, são acedidos por dia? | Nº de acessos de consulta aos dados |
| Obsolescência | Qual a frequência de actualizações aos dados? | Nº de operações de actualização por unidade de tempo |
| Oportunidade | Qual a percentagem dos dados que estão actualizados? | Nº de linhas actualizadas (por unidade de tempo) / nº total de linhas |
| | Qual a idade dos dados no sistema? | Percentagem de linhas superiores a determinada idade |

**Tabela 5-9 – Métricas para medir o grau de frescura dos dados.**

A operacionalidade na recolha dos valores referentes às avaliações dos dados pode assentar em estatísticas sobre os dados, em questionários respondidos pelos consumidores dos dados e na consulta aos ficheiros LOG das actividades do DW.

## Completude

A completude dos dados consiste na captura dos dados do mundo real necessários para a execução das actividades [Bobrowski et al., 1999]. Assim, aquando da tomada de decisão, os decisores não devem detectar ausências de valores. A ausência de valores pode ficar dever-se à indisponibilidade do SO em facultar mais dados e à execução dos processos de carregamento dos dados no DW provocar a inconsistência entre os dados. A forma de avaliação do cumprimento desta dimensão pode ser concretizada, de modo abstracto, como o quociente entre os dados devolvidos pelas respostas às consultas dos consumidores e os dados existentes no mundo real [Naumann & Rolker, 2000]. Algumas métricas sobre esta dimensão são apresentadas na tabela seguinte.

| Questão | Métrica |
|---|---|
| O atributo apresenta ausência de valor, mesmo quando de preenchimento obrigatório? | Percentagem das linhas que contêm valores ausentes em colunas de preenchimento obrigatório |
| As linhas da tabela de factos encontram-se carregadas nas dimensões? | Percentagem das linhas da tabela de factos que respeitam a integridade referencial |
| As linhas encontram-se estruturadas hierarquicamente nas dimensões? | Percentagem das linhas não estruturadas hierarquicamente nas dimensões |
| Quantas linhas foram carregadas com sucesso? | Percentagem dos registos carregados com sucesso |

**Tabela 5-10 – Métricas para avaliar a completude dos dados.**

Relativamente, às técnicas de recolha destas métricas pode-se recorrer a análises sobre os LOGs das actividades do DW, a aplicação de ferramentas de *data profiling* e a especificação de consultas directas sobre os dados.

## Interpretação

A interpretação consiste no facto dos dados se apresentarem em formato compreensivo e que facilita o seu entendimento. Esta dimensão tem associada, conforme referenciado anteriormente, uma forte raiz subjectiva. Por isso, em vista a obtenção duma avaliação mais concreta, procura-se avaliar alguns parâmetros relativos ao fornecimento de informações sobre os dados alvo de consulta (metadados). Esta situação faz prever, igualmente, o cumprimento de determinados pressupostos, principalmente, a formação dos consumidores e a adopção de terminologias aceites pela realidade organizacional. Assim, um modelo de medida capaz de aferir sobre a interpretação dos dados pode consistir no grau em que a informação divulgada respeita a capacidade técnica dos consumidores em manuseá-la [Naumann & Rolker, 2000] (tabela 5-11).

| Questão | Métrica |
|---|---|
| Os dados apresentados são facilmente interpretáveis? | Nº de elementos de informação indocumentados |

**Tabela 5-11 – Métrica para avaliação da interpretação dos dados.**

A auditoria aos dados e metadados pode representar o melhor modo de avaliação da interpretação. Complementarmente, justifica-se a adopção de um questionário, destinado aos consumidores, para avaliar esta dimensão.

## Relevância

A relevância respeita a circunstância dos dados se mostrarem úteis aquando da tomada de decisão. Os dados devem ser aplicáveis e úteis na concretização da tarefa em mãos [Pipino et al., 2002]. Este tema possui particular interesse, na medida que os dados disponibilizados pelos DWs devem ter aplicação efectiva, de outro modo, deixaria de fazer sentido possuir um DW composto por informação irrelevante (tabela 5-12). O povoamento dum DW com dados inúteis gera o aparecimento de dados dormentes, desnecessários para o cabal cumprimento das actividades. Estes dados são ainda responsáveis pelo desinteresse pelo sistema e na obstrução a um desempenho cabal das interrogações sobre o DW [Inmon et al., 1998].

| Questão | Métrica |
|---|---|
| Existem dados que não são acedidos? | Percentagem de tuplos ou colunas que nunca são resultado de respostas |

**Tabela 5-12 – Métrica de avaliação da relevância dos dados.**

A análise dos ficheiros LOG das actividades do DW, bem como, a existência dum repositório de metadados activo que mantenha o cadastro dos acessos aos dados mostra ser uma boa iniciativa na recolha de elementos relativos a esta métrica.

## Acessibilidade

A acessibilidade dos dados respeita à disponibilidade destes para consulta por parte dos consumidores [Vassiliadis, 2000]. Certamente, um DW pode mostrar-se inacessível por diferentes motivos, como sejam as falhas do sistema ou as naturais actualizações dos dados. Importa, minimizar a ocorrência das falhas inesperadas e reduzir a indisponibilidade por razões de actualizações aos dados (tabela 5-13).

| Questão | Métrica |
|---|---|
| Qual a disponibilidade do sistema? | Percentagem de tempo que o DWs se encontra inoperacional por falhas |
| Qual a disponibilidade transaccional? | Percentagem de tempo que o DWs se encontra inoperacional por actualizações dos dados |

**Tabela 5-13 – Métricas para a avaliação da acessibilidade dos dados.**

As informações dos ficheiros LOG sobre as actividades do DW permitem determinar as métricas de avaliação da acessibilidade dos dados.

# Capítulo 6

# Metricware – Avaliação dos Dados num SDW

## 6.1 Estudo de caso

### 6.1.1 Contextualização Prática

Tendo em vista mostrar a importância assumida pela gestão da qualidade dos dados, em ambientes de DW, efectuou-se o estudo de um caso específico sobre o tema em causa. O estudo de caso refere-se a uma organização da área do retalho de bens de consumo, que não será revelada por razões de confidencialidade. A organização em estudo dedica-se à comercialização de bens de consumo, em território nacional e internacional. Para isso, dispõe de locais de comércio, designados por unidades funcionais que, de acordo com a sua tipologia, escoam os bens aos consumidores. Os incrementos tecnológicos registados nas últimas décadas e as naturais exigências do mercado, conduziram a organização a apostar, adicionalmente, num sítio na *Internet*, como outro ponto de venda. De salientar, igualmente, o soberbo engrandecimento da organização ao longo das últimas duas décadas e possível de ser constatado pela taxa de crescimento das unidades funcionais e no aumento do volume de vendas do negócio.

Esta realidade conduziu à necessidade de implementação dum SDW, capaz de dar resposta e apoio às actividades comuns da gestão comercial. O SDW foi desenvolvido na segunda metade da década de noventa, com o objectivo central de suportar a tomada de decisão das actividades de negócio. Actualmente, o SDW encontra-se num estado de plena maturação, servindo de ferramenta de trabalho a um conjunto alargado de decisores. Logo, o sistema é assumido, pelos consumidores dos dados, como uma plataforma que sustenta o processo decisório. A experiência de

interacção dos intervenientes com o sistema proporcionou a obtenção de um histórico capaz de revelar, por um lado, algumas fragilidades existentes e por outro lado, uma gama de potencialidades do sistema a serem exploradas (e.g. introdução de processos de mineração dos dados). O foco de atenção do nosso trabalho consiste na análise dos dados do DM de vendas dos artigos por dia que sustenta as actividades decorrentes da gestão comercial da organização. Para tal, a organização disponibilizou uma amostra representativa dos dados do DM. O leque de decisões suportadas pelo DM constam desde decisões estratégicas até decisões operacionais, dependendo do agente de decisão em interacção com o sistema. Os utilizadores do DM são:

– Os directores de loja: no auxílio da gestão de *stocks*, aprovisionamento e reposição.

– Os directores comerciais: na gestão das gamas de artigos.

– Os gestores de topo, planeamento e controlo: na tomada de decisões estratégicas.

Neste contexto, prevê-se efectuar os necessários ajustamentos do sistema às necessidades dos consumidores e o consequente alinhamento estratégico com a organização. Assim, o capítulo irá apresentar uma breve descrição da arquitectura do SDW existente, mais especificamente, sobre o DM relativo às vendas dos artigos por dia; a enunciação dos processos envolventes ao manuseamento dos dados, em particular, aqueles relativos às operações de transformação ou introdução de padrões de qualidade nos dados; os problemas detectados durante os testes de medida dos índices de qualidade registados pelos dados e finalmente, a apresentação de algumas iniciativas de recomendação visando a resolução das anomalias detectadas.

### 6.1.2 Apresentação geral

A área de intervenção deste estudo de caso centra-se na gestão da qualidade dos dados em ambientes de DW, enquanto assunto determinante para o sucesso do sistema em causa e transversal a toda a estrutura envolvente à organização.

O dinamismo associado às inúmeras variáveis que influenciam directamente o processo de decisão, não se compadece com a manutenção de dados feridos na sua qualidade e pouco adequados às exigências dos consumidores. O reconhecimento dos dados como recurso estratégico fundamental e diferenciador no mercado concorrencial releva este tema como um assunto organizacional e estabelece, hoje em dia, os SDWs como uma plataforma tecnológica preponderante. Conhecer o percurso dos dados ao longo do SDW, nomeadamente, os processos envolvidos e os repositórios onde permanecem nas diferentes camadas da arquitectura do SDW, associada à

identificação das anomalias dos dados aí existentes, facilita a entendimento das causas e permite apontar estratégias de recomendação que os elevem aos patamares de qualidade desejados.

### 6.1.3 Motivação e objectivos do processo de análise

Ao longo desta dissertação tem-se pretendido relevar a problemática da qualidade dos dados em SDWs como um assunto organizacional. Deste modo, procurou-se focar os aspectos mais importantes em termos de quebras de qualidade ao nível dos dados existentes nos SDWs. Em consequência, tendo em vista debelar esses problemas nos dados, foi proposta uma plataforma de um sistema de gestão da qualidade dos dados em ambientes de DW e paralelamente uma área funcional responsável pela administração dos dados. Nesse sentido, a motivação para a realização deste estudo de caso consiste, precisamente, em abordar este assunto num contexto organizacional real. Reconhecendo os diversos temas explanados ao longo da dissertação no estudo de um caso concreto e apontando os caminhos mais apropriados para o enquadramento deste assunto nas organizações.

O objectivo geral deste trabalho consiste em estudar a problemática da qualidade dos dados e promover a melhoria da qualidade dos dados existentes do SDW organizacional. Assim, tendo em vista a concretização deste objectivo geral, são definidos os seguintes objectivos específicos:

- Reconhecer a qualidade dos dados como um assunto organizacional.
- Descrever os processos de transformação e limpeza dos dados.
- Confirmar a existência de factos (defeitos) sobre os dados.
- Estabelecer critérios sobre a qualidade dos dados.
- Obter índices de qualidade sobre os dados existentes.
- Classificar a natureza dos defeitos nos dados.
- Enunciar um conjunto de recomendações para solucionar os problemas diagnosticados.
- Promover a prevenção e a melhoria contínua dos dados do SDW, em geral e do DM de vendas, em particular.

### 6.1.4 O processo de análise

As iniciativas de melhoria dos níveis de qualidade dos dados no repositório resultam de duas ordens de factores. Por um lado, são determinadas pelo grau de maturidade ou consciência eviden-

ciada pela organização relativamente aos dados armazenados. Esta questão traduz-se em termos práticos pelo reconhecimento das preocupações com os dados e que deriva na existência de políticas e mecanismos agindo sobre estes. Por outro lado, exige um acto continuado de execução de procedimentos que visem a melhoria contínua dos dados e dos processos que os envolvem, ou seja, não se trata de um processo de uma única execução.

Nesse sentido, este trabalho encontra-se estruturado em quatro momentos distintos. O primeiro momento trata da descrição da arquitectura do SDW vigente sendo realizada a representação gráfica do processo de ETL dos dados no DM de vendas. Esta descrição será efectuada com base no modelo IPMAP e permite identificar visualmente a circulação dos dados pelas diferentes actividades envolvidas. Os resultados obtidos neste momento associados a um conjunto de entrevistas, reuniões e observações sobre a realidade existente na organização permitirá, num momento posterior, identificar o nível de maturidade da qualidade dos dados patenteada pela organização em causa. Esta identificação baseia-se nos princípios que norteiam o escalonamento sobre o amadurecimento das questões relativas à qualidade dos dados e que são propostos em [Adelman et al., 2005]. Num terceiro momento, será aplicado um conjunto de métricas (objectivas e subjectivas), previamente definidas e que actuam sobre os dados e os processos envolvidos. Os resultados obtidos pela aplicação das métricas possibilitam o seu manuseamento por modelos capazes de aferir sobre a qualidade existente nos dados e processos e analisar o cumprimento do nível de satisfação dos consumidores finais. Por condicionalismos de natureza temporal e logística organizacional, optou-se por um certo pragmatismo e objectividade nos testes a efectuar sobre os dados e nesse sentido, recorreu-se a instrumentos de análise dos dados (*data profiling*, limpeza e auditoria) que forneçam algumas medidas e estatísticas mais prementes dos dados da amostra facultada. Os resultados obtidos revelam um conjunto de factos possíveis de se converterem em assuntos sobre os dados, o que subscreve os princípios defendidos em [Olson, 2003] e [Wang, 1998]. O último momento, refere-se à enunciação de algumas iniciativas de incremento da qualidade dos dados. Estas iniciativas devem ser consideradas como recomendações visando a melhoria da qualidade dos dados e devem ser tidas em conta durante a fase de desenvolvimento do novo DM de vendas, como prova de conceito e uma vez aprovadas, aplicadas a todo o SDW.

## 6.2 Descrição do DM de vendas

### 6.2.1 Arquitectura do SDW

A descrição do SDW, seguidamente apresentada, resulta dum conjunto de entrevistas e reuniões efectuadas com os responsáveis pela manutenção operacional do DM de vendas dos artigos por

dia [Mendes, 2006]. De salientar que os entrevistados assumem, igualmente, a responsabilidade pelo desenvolvimento duma nova versão deste DM, que se encontra, actualmente, em fase de conclusão. O SDW da organização segue uma metodologia de implementação assente nos princípios da arquitectura de *bus* pura [Kimball et al., 1998]. Em especial, no que concerne à configuração do DW como a reunião dos diversos DMs que sustentam as actividades organizacionais. Actualmente, o volume do DW cifra-se em torno dos 4 TB de informação. O processo de povoamento dos DMs inicia-se pela recolha dos dados no SO, denominados por *front office*. O *front office* é responsável pela captação e armazenamento dos dados referentes às vendas diárias em cada uma das unidades funcionais (e.g. lojas). Posteriormente, estes dados são enviados para um outro local (ERP), responsável por manter as vendas das unidades funcionais agregadas ao dia. Os valores das diferentes unidades funcionais são vertidos numa única tabela composta pelas vendas agregadas. Esta tabela mantém os dados relativos às vendas de vários dias. As colunas da tabela de vendas são: artigo, fornecedor, unidade funcional, tipo de unidade funcional, documento, tipo de movimento (e.g. vendas líquidas, vendas brutas), movimento, moeda e valor. A justificação pela introdução deste novo componente resultou da necessidade de integração e conciliação entre os sistemas informáticos existentes nas unidades funcionais e do crescimento do volume de dados resultante da evolução e diversificação das actividades de negócio.

A introdução de novos dados referentes às actividades de negócio (e.g. artigos, fornecedores) compreende uma vasta e complexa gama de operações. Nomeadamente, a necessidade em complementar os dados inseridos, com os dados constantes no DW e em outros sistemas de dados, de modo a produzir linhas de valores que satisfaçam as diversas áreas de negócio e as exigências das diferentes aplicações (e.g. contabilidade). Neste contexto, os dados referentes aos artigos, fornecedores e unidades funcionais são mantidos por uma plataforma *Enterprise Application Integration* (EAI), do tipo *Enterprise Service Bus* (ESB), capaz de mostrar coordenadamente os dados e os processos organizacionais. Deste modo, procura-se conectar muitos sistemas independentes e disponibilizar uma única realidade sobre os dados da organização. De referir, que a solução ESB corresponde, segundo diversos estudos, à aproximação EAI tecnologicamente mais avançada. Cada aplicação publica, individualmente, mensagens no bus e subscreve, quando pretende receber mensagens do *bus*. Os dados das vendas residentes no EAI são relativos às unidades funcionais, aos fornecedores, aos movimentos e aos artigos. Assim, os sistemas ERP e EAI podem ser assumidos como as fontes de dados a integrar no DW.

A ARD acolhe os dados provenientes das fontes (ERP e EAI). O processo de circulação dos dados inicia-se com a transferência, para a ARD, da tabela de vendas agregadas, localizada no ERP. Esta tabela é particionada num número de tabelas igual à cardinalidade da tabela de vendas

agregadas (número de unidades funcionais). As colunas extraídas desta tabela são: artigo, fornecedor, unidade funcional, tipo de unidade funcional, documento, tipo de movimento, movimento e valor. Estes dados e os oriundos do EAI (artigos, unidades funcionais e fornecedores) são sujeitos a um escasso número de operações comuns à ARD, como sejam:

- O preenchimento com '0' no código da tabela fornecedor (tipo de dados não numérico).
- A concatenação do número da unidade funcional com a especificação do tipo de unidade funcional (e.g. L001).
- A eliminação dos espaços no final das cadeias de caracteres em algumas colunas.
- A inclusão de chaves de substituição.

Logo, as operações de transformação dos dados na ARD não contemplam procedimentos para:

- A estandardização e normalização de valores.
- O tratamento de valores nulos ou ausentes.
- A definição de valores por defeito.
- A imposição de regras de negócio.
- A identificação e reunião de valores duplicados.
- A correcção de valores.
- A preservação de consistência entre dados antigos e recentes.
- A averiguação de violações de integridade referencial.
- A falta de consistência dos dados.

A existência dum sistema EAI justifica a ausência destas operações, assumindo-se que os valores dos dados divulgados pelo EAI contemplam estas questões. Esta situação faz antever, igualmente, o diminuto consumo de tempo e recursos dispendidos nas operações de ETL e consequentemente, a existência duma janela de oportunidade suficientemente folgada para as exigências do negócio e que permite o refrescamento integral diário das dimensões em cerca de 30 minutos.

Na etapa final, os dados são carregados nos diversos DMs, portanto não existe um repositório do DW. Relativamente, ao DM actual registe-se o facto de se encontrar, presentemente, em fase de reestruturação em vista a simplificação do processo de povoamento dos dados, rectificação das deficiências detectadas (e.g. a implementação de chaves de substituição nas dimensões) e a me-

lhoria no desempenho no acesso aos dados. O DM de vendas inicial era composto pelas seguintes dimensões: *tempo*, *promoção* (dimensão degenerada), *unidade funcional*, *meio de divulgação*, *artigo* e *fornecedor*. A figura 6-1 ilustra simplificadamente a arquitectura do SDW.

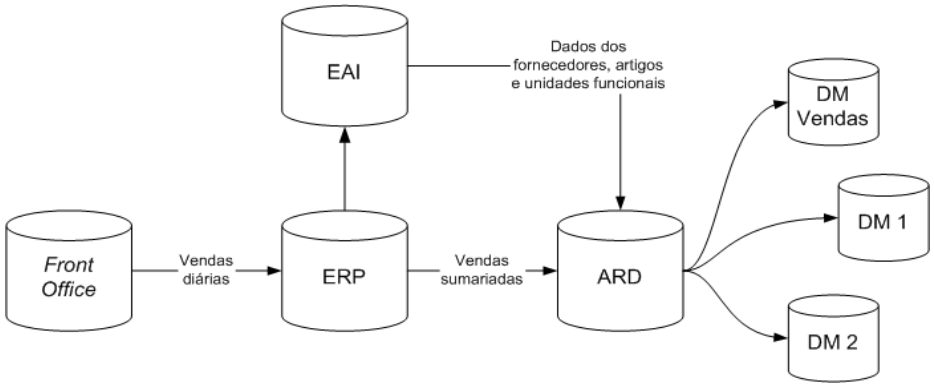



**Figura 6-1 – Arquitectura simplificada do SDW do estudo de caso.**

Durante a fase de transição do DM de vendas inicial para o novo DM de vendas, assinale-se a necessidade em manter algumas especificidades inerentes aos mesmos, como seja, a manutenção de uma dimensão artigo, que contempla chaves de substituição e uma outra, derivada da primeira, que não contempla chaves de substituição. O povoamento do DM de vendas novo é baseado e ocorre após o carregamento dos dados no DM de vendas antigo. De referir que o novo DM de vendas compreende a promoção dos artigos como uma tabela de factos sem medidas.

A taxa de crescimento médio diário do DW é cerca de 0,5%. Relativamente, à taxa de crescimento das dimensões, cifram-se os seguintes valores (em nº de linhas): *fornecedor*: 500/ano; *unidade funcional*: 150/ano; *artigo*: 500000/ano. Estes indicadores relevam a importância da dimensão *artigo*, uma vez que permite enquadrá-la como uma dimensão tendencialmente monstruosa [Kimball & Caserta, 2004]. A elevada taxa de crescimento desta dimensão deve-se ao surgimento de novos artigos, à diversificação dos ramos de negócio e ao ajustamento dos artigos em relação à categoria a que pertencem (e.g. um mesmo artigo pode frequentar diferentes categorias ao longo do seu ciclo de vida dentro da organização). A cardinalidade das dimensões: *artigo*, *unidade funcional*, *fornecedor* e *promoção* estima-se em 1 milhão e meio de linhas, 1100 linhas, 20 mil linhas e 19 mil linhas respectivamente. Enquanto, a tabela de factos é composta por 550 mil milhões linhas. Os DMs são a base para cerca de 100 mapas pré-formatados, OLAP e ferramentas *ad-hoc*. Os dados mais acedidos referem-se às dimensões: *unidade funcional*, *fornecedor* e *artigo*. Interessa ainda referir que a complexidade das consultas realizadas pelos utilizadores pode demorar desde escassos segundos até às horas e que o responsável pela qualidade dos dados no sistema é o administrador do DW.

### 6.2.2 O processo de ETL

Os processos inerentes às actividades de ETL resumem-se, praticamente, à extracção e carregamento dos dados nos DMs. As operações de transformação e limpeza dos dados realizadas constituem uma pequena parte das necessárias, de acordo com a qualidade dos dados observada. Em vista a descrição das tarefas envolventes ao processo de ETL, recorremos à aplicação do modelo IPMAP, como forma de representar esquematicamente as actividades envolvidas e a sua sequenciação [Shankaranarayan et al., 2000] [Shankaranarayan, 2005]. Seguidamente, é descrito o processo de ETL dos dados, desde as fontes até aos DMs.

#### Operações de extracção e limpeza dos dados

Os dados a extrair para a ARD provêm do EAI, com excepção dos dados relativos às vendas *on-line*, que resultam da extracção directa do *front office*. Os processos de limpeza dos dados existentes consistem, genericamente, na remoção dos espaços em branco nos valores de algumas colunas das tabelas artigo e fornecedor, no preenchimento da coluna referente ao código de fornecedor com valores iguais a zero, desde que o código existente não contenha cinco algarismos e na utilização de chaves de substituição. A figura 6-2 e representa as operações de extracção e limpeza dos dados extraídos do SO e as tabelas 6-1, 6-2, 6-3 e 6-4 descrevem a natureza das actividades envolvidas.

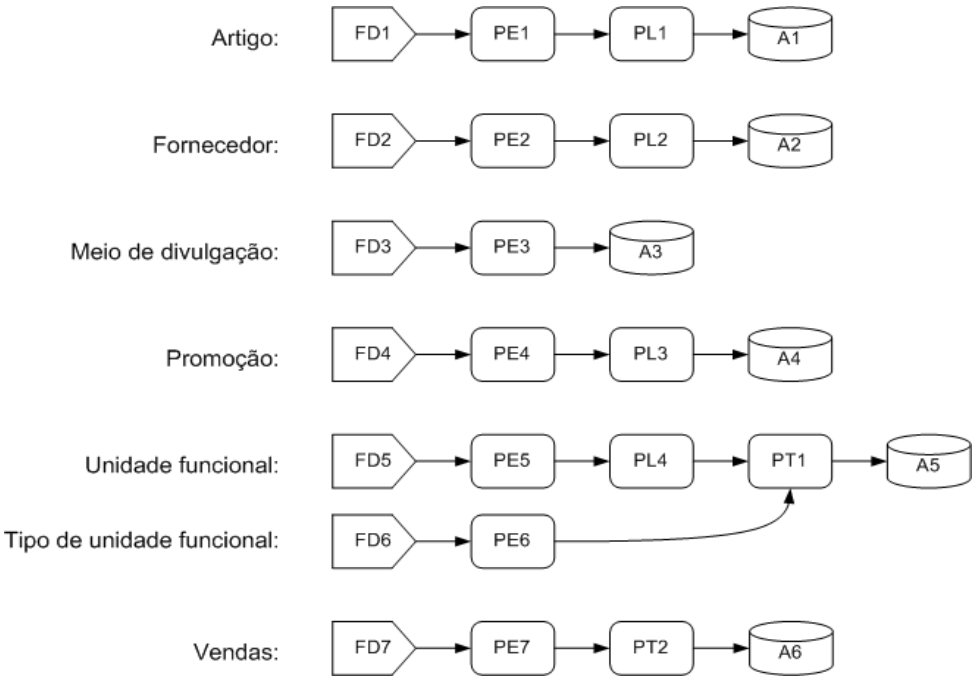



**Figura 6-2 – IPMAP relativo ao processo de extracção dos dados das fontes para a ARD.**

| Fonte de dados | Origem | Local | Destino |
|---|---|---|---|
| FD1 | Artigo | EAI | Artigo |
| FD2 | Fornecedor | EAI | Fornecedor |
| FD3 | Meio de divulgação | Interno | Meio de divulgação |
| FD4 | Promoção | EAI | Promoção |
| FD5 | Unidade funcional | EAI | Unidade funcional |
| FD6 | Tipo de unidade funcional | EAI | Tipo de Unidade funcional |
| FD7 | Vendas ERP (tabela única) | ERP | Vendas |

**Tabela 6-1 – Descrição das fontes de dados do SDW.**

| Processo | Origem | Destino | Política de circulação de dados |
|---|---|---|---|
| PE1 | Artigo | Artigo | Pull |
| PE2 | Fornecedor | Fornecedor | Pull |
| PE3 | Meio de divulgação | Meio de divulgação | Pull |
| PE4 | Promoção | Promoção | Pull |
| PE5 | Unidade funcional | Unidade funcional | Pull |
| PE6 | Tipo de unidade funcional | Unidade funcional | Pull |
| PE7 | Vendas ERP | Vendas | Pull |

**Tabela 6-2 – Descrição dos processos de extracção dos dados das fontes.**

| Processo | Objecto | Acção |
|---|---|---|
| PL1 | Artigo | – Remoção dos espaços das cadeias de caracteres.<br>– Conversão dos caracteres em maiúsculas.<br>– Aplicação de chaves de substituição. |
| PL2 | Fornecedor | – Remoção dos espaços das cadeias de caracteres.<br>– Conversão dos caracteres em maiúsculas.<br>– Preenchimento do código de fornecedor com zeros (campo *string*).<br>– Aplicação de chaves de substituição. |
| PL3 | Promoção | – Aplicação de chaves de substituição. |
| PL4 | Unidade funcional | – Aplicação de chaves de substituição. |

**Tabela 6-3 – Descrição dos processos de limpeza dos dados.**

| Armazenamento | Objecto |
|---|---|
| A1 | Artigo |
| A2 | Fornecedor |
| A3 | Meio de divulgação |
| A4 | Promoção |
| A5 | Unidade funcional |
| A6 | Vendas |
| A7 | Tabela de factos |
| A8 | Dimensão artigo |
| A9 | Dimensão fornecedor |
| A10 | Dimensão meio de divulgação |
| A11 | Dimensão promoção |
| A12 | Dimensão unidade funcional |

**Tabela 6-4 – Repositório de dados existentes.**

## Operações de transformação e carregamento dos dados no DW

As operações de transformação consideradas na tabela de factos consistem na simples substituição das chaves estrangeiras do SO pelas chaves de substituição criadas na ARD e o posterior carregamento dos dados no DW (figura 6-3) (tabela 6-5). Em relação às dimensões não são consideradas nenhumas operações de transformação prévias ao carregamento dos dados no DM (figura 6-4) (tabela 6-6), com excepção da concatenação do número de *unidade funcional* com o *tipo de loja funcional*, na tabela *unidade funcional* (figura 6-2).

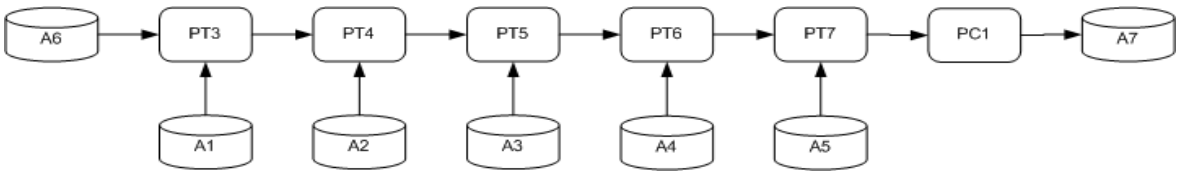



**Figura 6-3 – IPMAP dos processos de transformação e carregamento dos dados da tabela de factos.**

| Processo | Objecto | Acção |
|---|---|---|
| PT1 | Unidade funcional | Concatenação entre o tipo de unidade funcional e a unidade funcional (e.g. L001). |
| PT2 | Vendas | Partição da tabela de vendas sumariadas no número igual à cardinalidade da tabela. |
| PT3 | Vendas e artigo | Substituição das chaves do SO pelas chaves de substituição. |
| PT4 | Vendas e fornecedor | Substituição das chaves do SO pelas chaves de substituição. |
| PT5 | Vendas e meio de divulgação | Substituição das chaves do SO pelas chaves de substituição. |
| PT6 | Vendas e promoção | Substituição das chaves do SO pelas chaves de substituição. |
| PT7 | Vendas e unidade funcional | Substituição das chaves do SO pelas chaves de substituição. |

**Tabela 6-5 – Descrição dos processos de transformação e integração dos dados.**

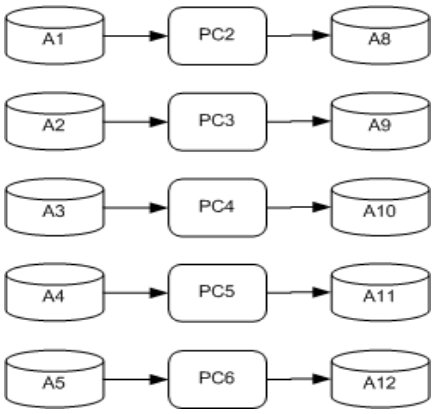



**Figura 6-4 – IPMAP relativo ao carregamento das dimensões no DM.**

| Processo | Objecto | Acção |
|---|---|---|
| PC1 | Vendas | Carregamento dos dados da tabela de factos no DM. |
| PC2 | Artigo | Carregamento dos dados da dimensão no DM. |
| PC3 | Fornecedor | Carregamento dos dados da dimensão no DM. |
| PC4 | Meio de divulgação | Carregamento dos dados da dimensão no DM. |
| PC5 | Promoção | Carregamento dos dados da dimensão no DM. |
| PC6 | Unidade funcional | Carregamento dos dados da dimensão no DM. |

**Tabela 6-6 – Descrição dos processos de carregamento dos dados no DM.**

## 6.3 Nível de maturidade da organização

De maneira geral, podemos considerar que o SDW da organização representa uma realidade muito particular no domínio deste tipo de sistemas. Esta observação resulta de um conjunto de questões, como sejam:

- O ritmo de evolução da organização e o espectro alargado das actividades do negócio.
- A idade do sistema e o consequente peso histórico dos dados e dos processos operativos.
- A rudimentar e escassa manutenção de metadados e documentação sobre os requisitos, processos e dados organizacionais.
- A complexidade envolvente aos fluxos circulatórios dos dados, de modo a facultar o acesso a diversas aplicações.
- As implementações sucessivas de componentes de *software* e *hardware* no sistema.
- A inexistência de um repositório do DW.
- O administrador do DW é o responsável máximo pelos dados.

Estes aspectos resultam, do ponto de vista da funcionalidade organizativa em inúmeras dificuldades de manutenção e melhoramento do DW sentidas pelos intervenientes de *back-end* do DW e, potencialmente, elevam as desconfianças dos utilizadores finais em relação às informações apresentadas. Porém, os responsáveis pelo sistema assumem uma postura de confiança relativamente aos dados disponibilizados (cf. [Mendes, 2006]). Situação que permite antever um desconhecimento sobre o estado real dos dados e que irá ser comprovada mais adiante neste trabalho.

Do ponto de vista do cumprimento dos objectivos previstos para a realização deste trabalho, as questões focadas podem provocar algumas contrariedades capazes de influenciar o percurso inicialmente traçado, sem contudo perverter o seu objectivo inicial. A primeira contrariedade resulta da dimensão organizacional e da complexidade do SDW que suporta as actividades de negócio porque impossibilita o conhecimento integral por parte dos interlocutores de toda a realidade existente. A segunda decorrente da anterior, respeita a razões logísticas, de índole temporal e organizacional, e que determinam o acesso a um menor nível de detalhe de algumas operações ou realidades sobre os dados e processos (e.g. IPMAP de alto nível). Por último, apesar do âmbito do trabalho centrar-se a nível do DM de vendas, revela-se pouco razoável realizar análises sobre o SO (e.g. avaliar a qualidade dos dados do EAI e do ERP). Apesar destas contrariedades a realidade observada permite-nos aferir sobre o nível de maturidade da organização relativamente à manutenção de dados de elevada qualidade.

Neste contexto, baseados nos critérios de escalonamento que avaliam o nível de maturidade das organizações, relativamente à qualidade dos dados, podemos afirmar que a realidade observada enquadra a organização no segundo nível, por ordem crescente, dos cinco níveis de maturidade propostos e que se designa por estado de despertar [Adelman et al., 2005]. A atribuição deste

nível de maturidade sobre a qualidade dos dados existentes na organização deve-se a diversos factores, como sejam:

- Algumas iniciativas prevêem a implementação de um programa visando a manutenção de dados de superior qualidade nos repositórios (e.g. processos de controlo entre aplicativos fonte e aplicativos destino).
- A inexistência de políticas pro-activas e preventivas para a garantia da qualidade dos dados, em particular, as que impeçam a introdução de dados irregulares (e.g. formação dos intervenientes, políticas de incentivos, auditoria e *profiling* dos dados, etc.).
- A escassa e rudimentar documentação e manutenção de metadados sobre os processos e dados.
- Alguns colaboradores pretendem incorporar iniciativas de disciplina de qualidade dos dados nos projectos a seu cargo.
- A ausência de responsáveis ao nível dos dados (administradores dos dados, administradores dos metadados e guardiães dos dados).
- A insuficiência de procedimentos de transformação e limpeza dos dados (e.g. tratamento de valores ausentes ou nulos).
- Inexistência de uma área funcional responsável pelos dados e orçamentada de verbas destinadas ao cumprimento das suas actividades.
- O responsável pelos dados é o administrador do DW.

Os responsáveis pelo sistema assumem a correcção dos dados, a inexistência de linhas sobrecarregadas e a estandardização de abreviaturas. É referido, igualmente, a existência de processos de controlo da qualidade que asseguram o avanço dos processos mensais e a satisfação dos critérios de qualidade predefinidos é realizada através de métricas que se encontram alinhadas com os aplicativos de onde é originária a informação [Mendes, 2006]. Aparentemente, parece estarmos em desacordo com os responsáveis, talvez, porque os critérios sejam cumpridos apenas sobre os dados mais críticos, como sejam os existentes na tabela de factos. Como na prática não temos a possibilidade de avaliar a qualidade dos dados neste repositório, temos de nos cingir somente aos dados da amostra e que não permitem cabalmente aferir sobre esta situação.

A nossa convicção é que a organização apesar de desperta para a problemática da qualidade dos dados, ainda não se encontra alerta para as graves consequências que a fraca qualidade dos dados acarreta e que consequentemente, se mostram impeditivas de futuras explorações mais

complexas sobre os dados. Esta realidade representa um cenário inicial de preocupação em relação aos dados e o presente estudo poderá contribuir como um possível ponto de partida tendo em vista encarar o assunto de maneira mais efectiva daquela vivida até ao presente momento. A implementação de uma gestão da qualidade dos dados do SDW e consequentemente, dos dados organizacionais, não é resultado de uma única execução e não pode ser conquistada por simples implementações de *software*. Antes, resulta de medidas concretas e continuadas de melhoria da qualidade dos dados e como tal, nunca será definitivamente solucionada. Ora, este processo de constante atenção sobre os dados, conduz a encará-los efectivamente como recursos estratégicos e a sabedoria na sua gestão determina o progresso da organização pelos patamares de maturidade definidos.

## 6.4 Problemas de qualidade nos dados

### 6.4.1 Indicadores de qualidade dos dados

Conforme referido anteriormente, a realidade vivida no seio da organização configura-a como pertencendo ao segundo patamar da escala de maturidade relativamente à qualidade dos dados. Assumindo, igualmente, que as iniciativas em torno dos dados não podem ser entendidas como um evento, mas antes como um processo dinâmico em contínua evolução, algumas modificações de análise mostram-se pertinentes no que respeita à avaliação desses dados. Preferencialmente, a aferição da qualidade dos dados proposta pretendia fazer uso de modos de avaliação subjectivos (e.g. respostas a questionários), métricas objectivas (e.g. interrogações directas aos dados) e modelos que facilitassem a interpretação desses valores e apontassem linhas de orientação nas iniciativas a seguir. Porém, estes instrumentos, apesar de importantes, mostram-se desajustados em relação ao estádio de maturidade verificado e assim, optou-se por uma atitude mais pragmática. O entendimento sobre o estado dos dados assentou na recolha de indicadores objectivos sobre os dados. Nesse sentido, efectuou-se a avaliação dos dados com recurso às potencialidades oferecidas pelas ferramentas de análise de dados (*data profiling*, limpeza e auditoria). Estas ferramentas são capazes de transmitir além de medidas objectivas e estatísticas sobre os dados, também conhecimentos sobre as características destes e que são possíveis de se consubstanciarem em metadados a integrar no sistema.

Os resultados obtidos pelo *software* considerado e os acessos directos aos dados servem de base na concretização do objectivo central proposto e que consiste em reconhecer a problemática da qualidade dos dados em ambientes de DW. Os resultados obtidos revelam a existência de factos sobre os dados (e.g. a coluna *cod_zona_preco*, da tabela *unidade funcional* apresenta 54% de

valores omissos). A junção destes factos pode revelar assuntos sobre os dados [Olson, 2003]. Os indicadores obtidos pela execução dos programas focam algumas dimensões concretas dos dados, em especial, aquelas que apresentam um condão intrínseco factual e independente das tarefas [Pipino et al., 2002]. Assim, são colhidas medidas sobre os dados relativamente:

- À ausência de valores nos dados.
- À relevância dos dados disponibilizados pelo sistema.
- À correcção dos valores armazenados em relação à realidade.
- À facilidade na interpretação dos dados divulgados.
- À consistência dos dados.

Conscientes da existência de outras medidas capazes de permitir a aferição sobre a qualidade dos dados, os critérios adoptados parecem os mais adequados e de possível averiguação, tendo em conta o nível de maturidade da organização e a concretização do objectivo proposto. Neste contexto, o estudo será realizado pela enunciação dos defeitos identificados nos dados, a sua catalogação e posterior indicação de resolução por um método adequado para o efeito.

### 6.4.2 Taxionomia das anomalias nos dados

A catalogação dos erros encontrados no DM de vendas assentou nas investigações desenvolvidas em [Kim et al., 2003], [Oliveira et al., 2005a, 2005b], [Adelman et al., 2005] e [Olson, 2003]. Em [Olson, 2003] é abordada a identificação das anomalias nos dados de modo conciso e pragmático, mas mostra-se demasiado simplista na tipificação dos erros existentes. Em [Adelman et al., 2005] são descritas mais detalhadamente as regras para a validação dos dados, em torno das dimensões mais representativas. Enquanto, as outras duas investigações abordam os defeitos nos dados baseando-se em refinamentos sucessivos de busca do problema alvo de observação. Nessas investigações, um problema dos dados é conduzido por patamares sucessivos de avaliação, de acordo com a sua tipologia, até ser encontrado o problema específico. As investigações consideradas encontram-se orientadas para a generalidade dos repositórios de dados e como tal, mostram-se exaustivas na identificação das anomalias verificadas nesses locais, sendo por isso, adequadas para repositórios de dados ainda não sujeitos a imposições de níveis de qualidade. Todavia, no âmbito dos DMs, muitas acções de melhoramento dos dados encontram-se geralmente já concretizadas (e.g. a adopção de chaves de substituição). Assim, iremos adoptar as terminologias consideradas válidas para o caso em estudo e adoptar outras que visem verificar a ocorrência de outros defeitos ao nível da qualidade dos dados, além dos considerados nestas investigações.

É o caso da relevância dos dados que se encontram ou não armazenados no repositório. Por outras palavras, podemos analisar a qualidade dos dados existentes em torno da qualidade do modelo multidimensional adoptado [Calero et al., 2001] e em termos do cumprimento dos requisitos dos utilizadores. Assim, importa aferir sobre a pertinência de determinadas colunas ou a ausência de outras passíveis de auxiliarem os decisores. Estas questões apesar de importantes não são alvo de análise na sua plenitude, especialmente, no que diz respeito à ausência de colunas, no modelo multidimensional, importantes para os consumidores finais (engenharia de requisitos). Tanto por razões de objecto de estudo, como de natureza logística, coloca-a num plano além do âmbito deste trabalho. Relativamente, à manutenção de colunas no modelo multidimensional, aparentemente, sem utilidade para as interrogações comuns por parte dos consumidores, iremos relacioná-las com a completude dos dados existentes. Apesar deste não ser um critério único, mostra-se uma condição necessária. O dinamismo associado ao processo de tomada de decisão condiciona as colunas a serem alvo de apreciação ou cálculo para indicadores. Por isso, importa abordar as colunas existentes que apresentam pouca probabilidade de acesso num futuro próximo. Um outro critério passível de ser avaliado é relativo à consistência do armazenamento dos valores de todas as colunas da tabela e que se pode relacionar com a interpretação dos dados, isto é, realizarem-se operações de transformação idênticas sobre os valores (e.g. uma coluna mantém valores 'Y' e 'N', enquanto outra alberga os valores 'S' e 'N'). Esta observação permite igualmente aferir por um lado, sobre a qualidade dos processos de transformação dos dados e por outro lado, sobre a natureza da qualidade do *software* produzido.

Em resultado das primeiras observações sobre os dados constantes no DM, verificam-se imperfeições nos dados comuns de ocorrer na ARD. Conforme referido anteriormente, a passagem dos dados, desde o SO até aos DMs, é realizada quase directamente. É comum designar-se esta realidade como "curto-circuito" do processo de DW. Neste contexto, interessa perceber o estado dos dados no DM alvo de estudo e antever a influência destes nas respostas às interrogações dos consumidores. Tendo em vista a classificação inequívoca das irregularidades verificadas nos dados constantes no DM em estudo, adoptamos uma taxionomia própria e fundamentada nas propostas enunciadas anteriormente. Esta taxionomia tem como objectivo facilitar o modo de resolução dos problemas encontrados (tabela 6-7).

| Nº | Anomalia | Descrição |
|---|---|---|
| 1 | Valor ausente | São os valores relevantes para os consumidores finais, mas que não se encontram presentes na estrutura multidimensional. |
| 1.1 | Linha incompleta | Consiste num significativo número de colunas, referentes a uma linha, que não se encontram devidamente preenchidas. |
| 1.2 | Coluna incompleta | Verifica-se quando um expressivo número de linhas, referentes a uma coluna, não se encontra preenchida. |
| 1.3 | Inexistência de coluna relevante | Corresponde à ausência de uma coluna considerada importante para a produção de resultados. |
| 1.4 | Inexistência de linha relevante | Corresponde à ausência de uma linha de dados considerada importante para a produção de resultados. |
| 2 | Valores irrelevantes | Refere-se à materialização de valores insignificantes (e.g. linhas ou colunas duplicadas). |
| 3 | Valores inválidos | São aqueles que violam um conjunto ou intervalo de valores possíveis. |
| 4 | Valores válidos | São os valores que contêm um conteúdo aceite pelo sistema. |
| 4.1 | Correctos | Apresentam o conteúdo com o valor correspondente ao mundo real. |
| 4.1.1 | Representação certa | São os valores que apresentam o conteúdo e o formato definidos pelo sistema. |
| 4.1.2 | Representação errada | São os valores que apresentam o conteúdo correcto, mas cujo formato se encontra desfasado do predefinido para o sistema. |
| 4.1.2.1 | Ambíguos | Corresponde aos valores que apresentam um conteúdo dúbio ou impreciso (e.g. meias desporto, meias futebol). |
| 4.1.2.2 | Incompletos | São os valores cujo conteúdo da coluna não se encontra totalmente preenchido (e.g. valores truncados). |
| 4.1.2.3 | Não estandardizados | São valores cujo conteúdo não está de acordo com o formato definido para a coluna (e.g. lda. e ltd). |
| 4.1.2.4 | Domínio inconsistente | Verifica-se quando não existe estandardização de conceitos entre colunas de conteúdo similar (e.g. numa coluna consta ‘y’ e ‘n’ e noutra consta ‘s’ e ‘n’). |
| 4.1.2.5 | Regras de negócio | São os valores que violam regras de negócio estabelecidas (e.g. 3 caracteres – 4 caracteres). |
| 4.1.2.6 | Desempenho | Os valores ocupam espaço desnecessário em disco. |
| 4.2 | Errados | Correspondem aos valores cujo conteúdo se apresenta incorrecto. |
| 4.2.1 | Erros ortográficos | Ocorre quando são inseridos erros ortográficos. |
| 4.2.2 | Incorrectos | Os valores que não correspondem à realidade (e.g. a moeda deixa de ser escudo e passa a euro). |
| 4.2.3 | Duplicação de valores | Ocorrência de diferentes valores para a mesma entidade. |
| 4.2.4 | Além do contexto da coluna | A coluna está sobrecarregada de valores além dos adequados para a coluna (e.g. inserção de símbolos). |
| 4.2.5 | Integridade referencial | Quando uma coluna que é chave estrangeira numa tabela contém um valor que não existe como chave primária na tabela relacionada. |
| 4.2.6 | Inconsistência de valores | Quando existem contradições entre os valores das colunas para uma mesma entidade. |
| 4.2.7 | Sem significado | Valores que apresentam um conteúdo com pouco ou nenhum significado. |

**Tabela 6-7 – Taxionomia de anomalias nos dados verificáveis em DMs.**

A resolução das anomalias prevê um conjunto de métodos de resolução dos defeitos. (tabela 6-8).

| Nº | Descrição |
|---|---|
| 1 | Decomposição dos dados em vista a obtenção de elementos atómicos. |
| 2 | Operações de transformação: |
| 2.1 | Estandardização (maiúsculas e minúsculas, acrónimos e abreviaturas). |
| 2.2 | Normalização (e.g. estabelecer regras de negócio). |
| 2.3 | Correcção. |
| 3 | Preenchimento de valores ausentes. |
| 4 | A aplicação das regras de integridade referencial. |
| 5 | Enriquecimento do conteúdo dos dados. |
| 6 | A resolução do problema dos valores duplicados nos dados das fontes. |
| 7 | Intervenção de perito. |
| 8 | Restringir a entrada de valores. |
| 9 | Reforçar a obrigatoriedade da entrada de valores. |
| 10 | Alterar tipo de dados. |
| 11 | Reformular o modelo multidimensional. |

**Tabela 6-8 – Métodos de resolução das anomalias nos dados.**

### 6.4.3 Sobre o software

As aplicações de software consideradas para a realização deste trabalho consistem em três tipos: *data profiling*, detecção de valores duplicados e auditoria aos dados (tabela 6-9). A justificação para a utilização destas aplicações consiste no facto da sua acção conjunta possibilitar uma recolha mais completa de informações relativas aos dados constantes no repositório. Nesse sentido, a aplicação de *data profiling* proporciona a captação de informações relativas ao conteúdo e estrutura da qualidade dos dados existentes nas tabelas. A ferramenta de detecção de linhas duplicadas procura, baseada num conjunto de parâmetros, a identificação de linhas similares sobre uma entidade, numa ou em várias tabelas. Neste caso, dado o caso em estudo incidir sobre um DM, será efectuada a pesquisa apenas sobre uma única tabela. Por fim, a aplicação de auditoria dos dados procura resolver os problemas nos dados, através do recurso a algoritmos de mineração de dados, capazes de revelar relacionamentos entre as diversas colunas da tabela. Adicionalmente, recorreu-se a consultas directas sobre os dados como forma de comprovar tendências ou problemas identificados pela utilização das aplicações.

| Aplicação | Tipo | Versão | Limitações | Site |
|---|---|---|---|---|
| *Datiris* | *Data profiling* | *Profiler Professional 1.2 – Trial* | 15 dias | www.datiris.com |
| *Wizsame* | Identificação de duplicados | *1.02 Demo* | Máximo 1000 linhas | www.wizsoft.com |
| *Wizrule* | Auditoria aos dados | *4.06 Demo* | Máximo 1000 linhas | www.wizsoft.com |

**Tabela 6-9 – Características do software utilizado no caso em estudo.**

A aplicação *datiris* é uma ferramenta de *data profiling* que possibilita a análise dos dados, através da capacidade em divulgar informações relativas aos dados existentes nas tabelas. As informações passíveis de serem obtidas relacionam-se com o conteúdo, a estrutura e a qualidade dos dados constantes no repositório. Estas informações possibilitam a ponderação sobre as prioridades e acções a tomar em vista debelar as deformidades ou problemas verificados. A recolha destas informações mostra-se ainda mais importante dada a inexistência de metadados referentes aos dados, processos e regras de negócio, no seio da organização em causa. Assim, as informações obtidas pela aplicação podem constituir-se como uma componente de metadados a integrar num plano superior de metadados do SDW.

A aplicação *wizsame* consiste num *software* de identificação de linhas semelhantes que representam a mesma entidade no mundo real. Após a detecção de linhas semelhantes poder-se-á proceder à fusão destas numa única linha, com as colunas relevantes e representativas da entidade em causa. No caso em estudo, recorremos a esta ferramenta no sentido de reforçar tendências de linhas apresentando valores duplicados e previamente detectadas pela aplicação de *data profiling* (cf. taxas de valores distintos sobre uma coluna).

Por último, a aplicação *wizrule* é uma ferramenta de auditoria dos dados, assente em tecnologia de mineração de dados. Este *software* revela as regras que relacionam os dados e aponta os casos de desvio às regras obtidas. As regras reveladas permitem aferir quanto a inconsistências, erros e casos alvo de auditoria.

Conscientes da existência de outras aplicações, certamente, dotadas de maiores potencialidades, importa referir que a opção pelo *software* utilizado se deve ao escasso número de ferramentas disponíveis em versões de teste e demonstrativas. É igualmente importante salientar que o recurso às aplicações consideradas consiste apenas como meio de trabalho na identificação de defeitos a nível da qualidade dos dados sobre o caso em estudo. Assim, apenas são expostas parte das capacidades patenteadas por estas aplicações, visto que o objectivo passa pela enunciação das anomalias dos dados e possíveis recomendações de tratamento e não a resolução dos problemas ou validação das ferramentas usadas. Conforme iremos constatar a ferramenta de audito-

ria foi pouco utilizada porque perante o panorama de qualidade dos dados verificado, as outras ferramentas mostraram-se mais adequadas e consequentemente produziram resultados mais objectivos, em especial, ao revelarem as características dos dados.

### 6.4.4 Esquema do DM de vendas

Os dados considerados, no caso em estudo, incidiram sobre o DM de vendas. Assim, com base no modelo multidimensional do DM, identificaram-se as seguintes dimensões: *tempo, artigo, fornecedor, promoção, meio de divulgação* e *unidade funcional* (figura 6-5). A dimensão *tempo* contém uma linha por cada dia do calendário. A dimensão *artigo* contém uma linha por cada artigo. A dimensão *fornecedor* retém uma linha por cada fornecedor. A dimensão *promoção* contém os dados relativos às promoções. A dimensão *meio de divulgação* detém uma linha por cada meio publicitário empregue nas campanhas publicitárias. Por fim, a dimensão *unidade funcional* retém uma linha por cada unidade funcional (loja, entreposto). Além das dimensões, consta igualmente a tabela de factos relativa às *vendas dos artigos*. Pelo facto do DM se encontrar num processo de reestruturação foram fornecidas ambas as tabelas de factos: a antiga, antes da reestruturação e a nova, resultante da reforma do DM. Porém, por questões de confidencialidade algumas colunas, críticas para a actividade do negócio, não foram facultadas pela organização. Esta restrição associada a uma pequena amostra das linhas disponibilizadas, inviabiliza uma avaliação sobre a tabela de factos. Tal situação não representa um entrave impeditivo à realização do trabalho proposto, apenas implica que alguns critérios de validação dos dados não sejam verificados. Assim, os dados analisados recaíram, essencialmente, nos valores constantes nas dimensões.

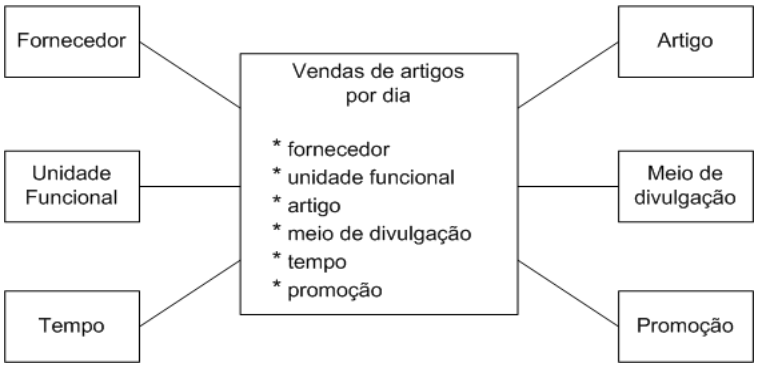



**Figura 6-5 – Esquema simplificado referente ao DM das vendas dos artigos.**

### 6.4.5 Algumas considerações gerais

A realização de uma análise mais objectiva sobre os dados pressupõe que sejam levadas em conta algumas considerações gerais. Relativamente às dimensões, procedeu-se a uma categorização segundo o modo de produção das linhas de dados. Nesse sentido, consideraram-se as dimensões que contêm dados provenientes do SO e aquelas que resultam de processos internos ao SDW. Nesta última categoria, os dados são gerados por processos precisos e automáticos, não estando sujeitos a intervenções humanas e operações comuns sobre os dados (e.g. actualização de valores). Por este facto, não consideramos estes dados como prioritários em termos de avaliação da qualidade e por isso, não são alvo de análise. Desta categoria fazem parte as dimensões *tempo* e *meio de divulgação*. Assim, as dimensões consideradas possíveis para avaliação, no âmbito deste trabalho, são: *fornecedor*, *promoção, artigo* e *unidade funcional*. As características gerais das dimensões podem ser observadas na tabela 6-10.

| Vista | Tamanho (em linhas) | Linhas activas (*status_key*='A') | | Linhas inactivas (*status_key*='I') | | Nº colunas |
|---|---|---|---|---|---|---|
| Artigo | 1 434 126 | 852 659 | *(59,5%)* | 581 467 | *(40,5%)* | 51 |
| Unidade funcional | 1 183 | 897 | *(75,8%)* | 286 | *(24,2%)* | 46 |
| Promoção | 19 154 | 19 154 | *(100%)* | - | *(0%)* | 13 |
| Fornecedor | 33 442 | 15 901 | *(47,5%)* | 17 541 | *(52,5%)* | 18 |

**Tabela 6-10 – Propriedades gerais das tabelas a analisar.**

Por imposição das ferramentas utilizadas (e.g. limite de linhas a avaliar) e do volume de dados a estudar, o modo de abordagem, em termos de análise, mereceu necessariamente uma devida reflexão. Sempre que possível as primeiras análises efectuadas reflectiram-se sobre a totalidade das linhas das tabelas. Os resultados obtidos servem de ponto de partida para análises mais específicas, mais manuseáveis e capazes de melhor revelarem anomalias nos dados. A definição de análises mais específicas efectuou-se segundo dois critérios: as colunas apresentarem o campo *status_key* igual a 'A' e sobre tranches de mil linhas. A justificação para o estabelecimento de lotes de mil linhas deve-se ao facto das ferramentas consideradas serem versões limitadas nas suas capacidades de utilização. De qualquer modo, esta situação é perfeitamente possível à luz de [Kimball & Caserta, 2004], porque inclusivamente faculta uma maior flexibilidade dos dados a examinar. O propósito que residiu à definição das vistas com a coluna *status_key* igual a 'A' consistiu no facto de considerar estas linhas como as activas (maior probabilidade de interrogações) e mais recentes e por isso, supostamente, as que apresentam maiores índices de qualidade, em resultado de operações de tratamento e limpeza mais efectivas sobre os dados. Esta constatação deriva de uma maior homogeneização dos valores dos dados registados (e.g. apenas maiúsculas, sem linhas duplicadas, sem espaços em branco antes ou após os valores dos dados e sem carac-

teres especiais). Adicionalmente, é possível observar a existência de duas colunas, *data_activacao* e *data_inactivacao*, que balizam o período durante o qual as linhas se encontram activas no sistema. Verifica-se que esta abordagem sobre a manutenção das linhas nas dimensões segue a aproximação de tipo 2 para as dimensões de variação lenta [Kimball et al., 1998]. As linhas com a coluna *status_key* igual a 'I', apresentam a coluna *data_inactivacao* com o preenchimento duma data já passada. Esta situação induz que se tratam de linhas actualmente inactivas, mas que são mantidas no sistema de modo a preservar a história e as interrogações temporais dos dados. Saliente-se o peso destes dados representar nas dimensões *artigo*, *fornecedor* e *unidade funcional*, cerca de 41%, 53% e 24% respectivamente do volume total de linhas.

A opção por estas vistas releva, igualmente, as discrepâncias estruturais na manutenção de dados frescos e antigos no mesmo repositório de dados e sugere dois vectores de incumprimento das características basilares sobre a qualidade dos dados. Primeiro, a possibilidade de incorrecções, incumprimento de políticas de normalização e tratamento dos dados, implicando certamente, incongruências nos dados divulgados aos consumidores finais. Segundo, o facto do DM de vendas albergar dados dormentes, nunca ou raramente utilizados [Inmon et al., 1998], o que implica o desperdício de espaço em disco, potencia a perda de desempenho no acesso aos dados e eficácia no tratamento destes. Esta constatação pode ser perspectivada em termos da manutenção de políticas eficazes de tendências sobre a utilização dos dados (metadados) e consequente remoção dos dados que manifestamente não mostrem indícios de utilização. Interessa salientar que a remoção destes dados do DM ou do repositório que os consumidores acedem directamente, não deve ser confundida com a sua remoção do sistema. Assim, os dados removidos podem e devem permanecer para eventuais interrogações futuras num local mais apropriado e que contem os dados menos utilizados nas consultas quotidianas e de maior nível de detalhe (ODS/DW).

Numa primeira análise sobre a composição das diferentes dimensões ressaltam algumas observações. A primeira consiste no elevado número de linhas registado pela dimensão *artigo*, que se justifica pela expansão da organização e pelas características inerentes à área de negócio em que se insere. O ritmo de crescimento desta tabela mostra que se trata de uma dimensão tendencialmente monstruosa [Kimball et al., 1998] [Kimball & Caserta, 2004]. Uma segunda observação baseia-se no significativo número de colunas apresentado pelas dimensões *artigo* e *unidade funcional*. Esta auscultação indicia a necessidade em manter os valores das colunas preenchidos de modo a responderem perante os acessos dos consumidores finais. A questão mostra-se ainda mais pertinente em relação à dimensão *artigo*, devido ao número de linhas existente e consequente volume de ocupação de espaço em disco. A justificação para o elevado número de colunas resulta da satisfação de requisitos pontuais e específicos dos consumidores, num dado instante

temporal. Na prática, dada a inexistência de metadados que registem os perfis de acesso e taxas de utilização dos dados, os valores destes dados não se encontram devidamente mantidos, originando além de inúmeras inconsistências, também a desactualização e ausência de muitos valores. Uma última observação prende-se com o expressivo número de linhas que se mantêm nas dimensões, mesmo estando inactivas. Este reparo pretende salientar que estas linhas apresentam uma tendência de baixa probabilidade de acesso.

Dado o elevado número de linhas existente na dimensão *artigo* e conforme foi dado a observar numa análise geral à tabela, optamos por deixar de incluir esta tabela no estudo. A fundamentação para esta posição baseia-se, em primeiro lugar, na verificação da permanência da generalidade dos problemas detectados nas outras dimensões. Assim, a escolha recaiu sobre dimensões que de algum modo representam realidades distintas entre si. A dimensão *fornecedor* mostra-se a tabela que apresenta, objectivamente, os melhores índices de qualidade nos dados. A dimensão *unidade funcional*, apesar de conter o menor número de linhas, continua a manifestar enormes anomalias nos dados. A dimensão *promoção* apresenta a coluna *status_key*, em todas as linhas, igual a 'A', contudo denota além dos defeitos encontrados nas outras dimensões, também uma alta propensão para a duplicação de linhas. Quanto à dimensão *artigo*, a única deformidade detectada, não verificável nas outras dimensões, é o facto das colunas *manutencao_geral* e *is_sale* assumirem valores 'Y' e 'N' como pertencentes ao domínio. Esta particularidade aponta para a possibilidade de incrementos progressivos ao sistema existente, realizados por diferentes equipas de programação e sem o cuidado de respeitar algumas características constantes do próprio sistema. Esta situação torna-se ainda mais delicada dada a ausência de metadados sobre os dados, esquemas e processos envolvidos. Uma outra razão que justifica a ausência da dimensão *artigo* deste estudo reside na pouca flexibilidade em termos de manuseamento dos dados pelas diversas ferramentas. Mesmo considerando válida a construção de lotes de mil linhas, estes podem apresentar-se pouco representativos da realidade dos dados contidos na tabela. Conforme referido anteriormente, os dados constantes no DM sofrem ligeiras transformações face aos existentes no SO, por isso, são detectadas falhas básicas na qualidade dos dados, capazes de desvirtuarem o lote de dados em análise. Por exemplo, uma consulta sobre as primeiras mil linhas da tabela, com a coluna *status_key* igual a 'A' e ordenada ascendentemente pela coluna *descr_artigo*, retorna apenas linhas que contêm o primeiro carácter igual a '#' nesta coluna.

### 6.4.6 Análise dos dados

## Dimensão unidade funcional

A análise dos dados desta tabela foi realizada sobre quatro vistas materializadas (tabela 6-11). A primeira, é sobre a totalidade da tabela, uma vez que o número de linhas é reduzido (1183 linhas) e por isso, torna-se facilmente manuseável em termos de análise. A segunda é sobre as linhas que contêm a coluna status_key='A'. Enquanto, a terceira vista considera as linhas que apresentam a coluna status_key='I'. A última vista contém as primeiras 999 linhas da tabela original e surge devido às limitações das aplicações *wizsame* e *wizrule*.

| Nº | Vista | Tamanho (em linhas) | Descrição | Status_key |
|---|---|---|---|---|
| 1 | Unidade_funcional | 1183 | Todas as linhas e colunas da tabela original. | - |
| 2 | Unidade_funcional_a | 897 | As linhas com a coluna *status_key* = 'A'. | A |
| 3 | Unidade_funcional_i | 286 | As linhas com a coluna *status_key* = 'I'. | I |
| 4 | Unidade_funcional_999 | 999 | As primeiras 999 linhas da tabela original. | - |

**Tabela 6-11 – Vistas definidas para a dimensão *unidade funcional*.**

A análise de *data profiling* sobre a vista 1 revela elevados índices de valores ausentes num número significativo de colunas, encontrando-se muitas linhas semi-vazias (tabela 6-12). Cerca de 50% das colunas apresentam elevados índices de valores nulos, zeros ou brancos. Destas, cerca de metade revelam taxas de ausência acima dos 50%. Se analisarmos detalhadamente a vista sobre as linhas mais recentes (vista 2) aponta para taxas de ausência de valores nulos e brancos, muito superiores (tabela 6-13). Cerca de dois terços das colunas mostram taxas de ausência superiores a 50%. Dado que a incidência de valores nulos, zeros e brancos se agrava na vista que contem os registos mais recentes, podemos deduzir que estas colunas não se mostram críticas para a organização e consequentemente tendem, no presente e no futuro, a não serem alvo de interrogações pelos consumidores. Além desta observação é possível constatar que determinadas colunas contêm valores válidos, mas incorrectos e que inviabilizam qualquer tipo de análise. É o caso da coluna *area_venda* que apresenta o mesmo valor '1' em 92 % das linhas. Logo, podemos considerar estes dados como dormentes porque aparentam uma certa irrelevância ou inutilidade quando colocados no DM afim de serem alvo de interrogações. Assim, consideramos inapropriada a existência destas colunas no DM, devendo antes preservá-las num local (ODS/DW) que contém os dados armazenados num maior nível de detalhe e passível de realizar melhorias da qualidade. Relativamente, aos valores por defeito predefinidos encontram-se diferenças de representação. Alguns dos valores por defeito utilizados são: ? e ???. Acresce que os valores por defeito atribuídos representam uma minoria face aos valores nulos, brancos e zeros.

| Coluna | Distintos | | Nulos | Zeros | Brancos | Padrões | Observações |
|---|---|---|---|---|---|---|---|
| Cod_und_funcional | 897 | 76% | | | | | |
| Descr_und_funcional | 948 | 80% | | | | 554 | |
| Sigla_8 | | | 1,4% | | 0,5% | 60 | 8 caracteres |
| Área_venda | | | | 3,8% | | | 1 = 1106 (93,5%) |
| Descr_zona | | | 36% | | | | ? = 121 |
| Primeiro_ano | | | | 43% | | | 0 = 507 |
| Primeiro_mes | | | | 43% | | | 0 = 507 |
| Primeiro_ano_cc | | | | 57% | | | 0 = 674 |
| Primeiro_mes_cc | | | | 57% | | | 0 = 674 |
| Descr_cc | | | 53% | | | 148 | |
| Insígnia_cc | | | 57% | | | | |
| Data_abertura_cc | | | 57% | | | | |
| Descr_ins_cc | | | 57% | | | | |
| Cod_zona_dop | | | 43% | | | | ? = 71 |
| Desc_zona_dop | | | 43% | | | | ? = 71 |
| Cod_zona_preco | | | 54% | | | | ? = 156 |
| Descr_zona_preco | | | 54% | | | | ? = 156 |
| Cod_zona_regiao | | | 38% | | | | ? = 37 |
| Descr_zona_regiao | | | 38% | | | | ? = 39 |
| Cod_fornecedor | | | 33% | | | | |
| Und_funcional_mig | | | 34% | | | | ? = 33 |
| Cod_cadeia | | | 32% | | | | ? = 31 |
| Descr_cadeia | | | 32% | | | | ? = 31 |
| Cod_zona_cluster | | | 63% | | | | ? = 170 |
| Descr_zona_cluster | | | 63% | | | | ? = 170 |
| Data _fecho | | | 76% | | | | |

**Tabela 6-12 – Características dos dados da vista 1 da dimensão *unidade funcional* (*Datiris*).**

| Coluna | Distintos | | Nulos | Zeros | Brancos | Padrões | Observações |
|---|---|---|---|---|---|---|---|
| Cod_und_funcional | | 100% | | | | | |
| Descr_und_funcional | 895 | 99,8% | | | | 492 | |
| Sigla_8 | 831 | 93% | 2% | | 0,7% | 59 | 8 caracteres |
| Área_venda | | | | 5% | | | 1 = 826 (92%) |
| Descr_zona | | | 46% | | | | Sem valor por defeito |
| Primeiro_ano | | | | 51% | | | 0 = 456 |
| Primeiro_mes | | | | 51% | | | 0 = 456 |
| Primeiro_ano_cc | | | | 60% | | | 0 = 534 |
| Primeiro_mes_cc | | | | 60% | | | 0 = 534 |
| Descr_cc | | | 55% | | | | |
| Insígnia_cc | | | 59% | | | | |
| Data_abertura_cc | | | 59% | | | | |
| Descr_ins_cc | | | 59% | | | | |
| Cod_zona_dop | | | 56% | | | | Sem valor por defeito |
| Desc_zona_dop | | | 56% | | | | Sem valor por defeito |
| Cod_zona_preco | | | 69% | | | | Sem valor por defeito |
| Descr_zona_preco | | | 69% | | | | Sem valor por defeito |
| Cod_zona_regiao | | | 50% | | | | Sem valor por defeito |
| Descr_zona_regiao | | | 50% | | | | Sem valor por defeito |
| Cod_fornecedor | | | 44% | | | | 00000 = 11 |
| Und_funcional_mig | | | 45% | | | | Só tem 'S' |
| Cod_cadeia | | | 42% | | | | Sem valor por defeito |
| Descr_cadeia | | | 42% | | | | Sem valor por defeito |
| Cod_zona_cluster | | | 80% | | | | Sem valor por defeito |
| Descr_zona_cluster | | | 80% | | | | Sem valor por defeito |
| Data _fecho | | | 98% | | | | |

**Tabela 6-13 – Características dos dados da vista 2 da dimensão *unidade funcional* (*Datiris*).**

Após a execução da aplicação de *data profiling*, procedemos à enunciação, pormenorizada, das imperfeições detectadas nos dados desta tabela (tabela 6-14), a sua catalogação e o método de resolução. A análise é realizada sobre a vista 2 porque apresenta maior homogeneidade entre os valores dos dados e se apresenta livre de linhas duplicadas.

| Coluna | Descrição | Problema | Valor original | Método | Valor transformado |
|---|---|---|---|---|---|
| Area_venda | Presença do mesmo valor num significativo número de linhas | 4.2.2 | 1 | 7 | |
| Centro_comercial | Valores nulos: ? ??? espaço | 4.1.2.3 | | 2.1 | |
| Cod_zona<br>Cod_zona_dop<br>Cod_zona_preco<br>Cod_zona_regiao<br>Cod_cadeia<br>Cod_zona_cluster | Colunas com valores idênticos à sua descrição. | 2 | ZNT *(NORTE)*<br>BAT *(BATATAS)*<br>LSB *(LISBOA)*<br>CNT *(CENTRO)*<br>BATATAS *(BATATAS Hipermercados)*<br>D *(CLUSTER D)* | 11<br>7 | |
| Cod_funcionario | ? espaço 00000 | 4.1.2.3 | | 2.1 | |
| Cod_zona<br>Descr_zona<br>Descr_cc | Presença de inconsistências entre colunas. | 4.2.6<br>4.2.3 | Z; NORTE<br>Z; LISBOA<br>ZNT; NORTE; MCC-Sintra | 4<br>7<br>2.3 | ZNT; NORTE<br>ZLX; LISBOA<br>ZLX; LISBOA; MCC – SINTRA |
| Desc_und_funcional | Erros ortográficos | 4.2.1 | MCC Angra d Heroísmo<br>MCC portimao<br>CWM SA Coimbra Stadm | 2.3 | MCC ANGRA DO HEROÍSMO<br>MCC PORTIMÃO<br>CWM SA COIMBRA STADIUM |
| Descr_und_funcional<br>Descr_zona_dop<br>Descr_cc | Valores em maiúsculas e minúsculas. | 4.1.2.3 | CVL Amadora<br>BOM CAFFE GUIA | 2.1 | CVL AMADORA<br>BOM CAFFE GUIA |
| Descr_und_funcional | Valores omissos ou incompletos. | 4.1.2.2 | GAZELA<br>GAZELA ALVERCA | 2.1<br>9 | GAZELA AVEIRO<br>GAZELA ALVERCA |
| Descr_und_funcional | Valores ambíguos (duplicados) | 4.1.2.1 | LOJA LEÃO G1<br>LOJA LEÃO G1 | 5 | LOJA LEÃO G1 – NRT<br>LOJA LEÃO G1 – SUL |
| Descr_und_funcional | Sem estandardização / normalização de valores e conceitos. | 4.1.2.3 | MADRID (MUITO SOL)<br>MS-VALENCIA<br>MUS Oeiras<br>MUITO SOL ALVERCA | 2.1<br>2.2 | MUITO SOL MADRID<br>MUITO SOL VALENCIA<br>MUITO SOL OEIRAS<br>MUITO SOL ALVERCA |

**Tabela 6-14 – Defeitos dos valores dos dados sobre a vista 2 da dimensão *unidade funcional* (continua).**

| Coluna | Descrição | Problema | Valor original | Método | Valor transformado |
|---|---|---|---|---|---|
| Descr_und_funcional<br>Formato<br>Universo<br>Centro_comercial<br>Descr_empresa<br>Descr_insignia<br>Cod_zona<br>Descr_zona<br>Descr_cc<br>Universo_cc<br>Descr_ins_cc<br>Descr_zona_dop<br>Descr_zona_preco<br>Descr_zona_regiao<br>Descr_zona_cluster | Tem espaços em branco após o valor. | 4.2.1. | POTATO MASTERS_____<br>GCT ON LINE _____ | 2.3 | POTATO MASTERS<br>GCT ON LINE |
| Descr_und_funcional | Contém valores além do contexto da coluna. | 4.2.4 | MCC9707 (MACACO)<br>EXP 2002 OL.HOSPITAL<br>CAV_EXP05 | 1 | |
| Primeiro_ano | Violação do intervalo. | 4.2.2 | 2, 3, 0, 2005 | 8 | |
| Sigla_8 | Representa uma abreviatura ou acrónimo da descr_und_funcional?<br>Quais as regras de negócio que estabelecem a atribuição de valores nesta coluna?<br>Apresenta cerca de 60 padrões distintos numa coluna com 8 caracteres (cf. resultados de *data profiling*). | 4.1.2.3 | MC-CACEM<br>MES-VISE<br>MESA CAB (Leiria)<br>MS-BRAGA<br>467-VASC | 1<br>2.1<br>2.2 | MC-CACEM<br>MC-VISEU<br>MC-LEIRIA<br>MC-BRAGA<br>MC-VASCO |

**Tabela 6-14 – Defeitos dos valores dos dados sobre a vista 2 da dimensão *unidade funcional* (continua).**

| Coluna | Descrição | Problema | Valor original | Método | Valor transformado |
|---|---|---|---|---|---|
| Descr_zona<br>Descr_cc<br>Insígnia_cc<br>Data_abertura<br>Descr_ins_cc<br>Cod_zona_dop<br>Descr_zona_dop<br>Empresa_mig<br>Cod_zona_preco<br>Descr_zona_preco<br>Cod_zona_regiao<br>Descr_zona_regiao<br>Cod_fornecedor<br>Und_funcional_mig<br>Cod_cadeia<br>Descr_cadeia<br>Cod_zona_cluster<br>Descr_zona_cluster<br>Data_fecho | Presença significativa de valores nulos nas colunas (cf. resultados de *data profiling)*.<br>Aparente indistinção entre valores ausentes e nulos.<br>Linhas semi-vazias.<br>Sem valores por defeito (e.g. CENTRO). | 1.2<br>1.1 | Espaço ? | 9<br>3<br>11 | ? |
| Primeiro_ano<br>Primeiro_mes<br>Primeiro_ano_cc<br>Primeiro_mes_cc | Presença expressiva de valores iguais a zero. | 1.2 | - | 11 | - |
| Unidade_funcional_mig | Além da ausência de valores, só tem valores 'S' | 2 | | 11 | |

**Tabela 6-14 – Defeitos dos valores dos dados sobre a vista 2 da dimensão *unidade funcional* (continuação).**

## Dimensão fornecedor

A análise dos dados da dimensão *fornecedor* foi realizada com base na materialização de quatro vistas (tabela 6-15). A primeira corresponde à totalidade da tabela e serve de repositório base para a aplicação *Datiris*. A segunda e terceira vista são estabelecidas sobre as linhas que se apresentam activas e inactivas, respectivamente. A última vista é relativa às primeiras 999 linhas da tabela original, com a coluna *status_key*='A' e ordenada ascendentemente pela coluna *razao*. Esta vista deve-se às limitações das aplicações *wizsame* e *wizrule*.

| Nº | Vista | Tamanho (em linhas) | Descrição | Status_key |
|---|---|---|---|---|
| 1 | Fornecedor | 33 442 | Todas as linhas e colunas da tabela original. | - |
| 2 | Fornecedor_oa_a | 15 901 | As linhas com a coluna *status_key* = 'A'. | A |
| 3 | Fornecedor_oa_i | 17 541 | As linhas com a coluna *status_key* = 'I'. | I |
| 4 | Fornecedor_oa_a_999 | 999 | Contém as primeiras 999 linhas da tabela com a coluna *status_key*='A' e ordenada ascendentemente pela coluna *razao*. | A |

**Tabela 6-15 – Vistas definidas para avaliar a qualidade dos dados da dimensão *fornecedor*.**

Uma primeira análise sobre a dimensão *fornecedor* permite considerar esta tabela como a que aparenta os melhores índices de qualidade dos dados do universo das tabelas alvo de análise. Esta constatação pode ser explicada pelo facto de ser a única que interage com o ambiente externo à organização e por isso, ser merecedora de maiores preocupações quanto a padrões e requisitos de qualidade, de modo a facilitar a identificação inequívoca dos fornecedores (tabela 6-16).

| Coluna | Distintos | | Nulos | Zeros | Brancos | Padrões | Observações |
|---|---|---|---|---|---|---|---|
| Cod_fornecedor | 15 901 | 48% | | | | | Existe o fornecedor zero? |
| Razão | 13 674 | 41% | | | | 9 986 | |
| Nome_comercial | 15 955 | 48% | | | | 6 903 | |
| Aecn | | | | | | | 999999 = 33 440 |
| Fornecedor_p | | | | | | | 00000 = 27 876; ? = 271 |

**Tabela 6-16– Características dos dados sobre a vista 1, da dimensão *fornecedor* (*Datiris*).**

Uma análise sobre os resultados obtidos pelo processo de *data profiling*, considerando a vista 2, permite confirmar a melhoria do nível de qualidade dos dados referentes à dimensão *fornecedor*, face às restantes dimensões (tabela 6-17).

| Coluna | Distintos | | Nulos | Zeros | Brancos | Padrões | Observações |
|---|---|---|---|---|---|---|---|
| Cod_fornecedor | | 100% | | | | | |
| Razão | 13 627 | 86% | | | | 9 952 | |
| Nome_comercial | 15 775 | 99% | | | | 6 672 | |
| Aecn | | | | | | | 999999 = 15 900 |
| Fornecedor_p | | | | 83% | | | 00000 = 13 244; ? = 221 |
| Fornecedor_modis | | | | | | | N = 15 855 S = 46 |

**Tabela 6-17– Características dos dados sobre a vista 2, da dimensão *fornecedor* (*Datiris*).**

Uma análise directa sobre a tabela leva-nos a considerar que apresenta a melhor qualidade dos dados. Efectivamente, nenhuma das colunas apresenta valores nulos, zeros ou brancos, situação que se traduz, provavelmente, por uma maior necessidade em manter os valores dos dados com garantias de maior qualidade. Mais especificamente, observamos os valores das colunas *razao* e *nome_comercial* em maiúsculas, bem como, a inexistência de espaços em branco antes e após os valores armazenados. A melhoria verificada nos dados será resultado da aplicação de algumas rotinas de transformação e limpeza dos dados. A excepção, a esta acentuada melhoria da qualidade dos dados, verifica-se na coluna *fornecedor_p* que contém o código '00000' em 83% das linhas e também 271 linhas com valor '?' (de qualquer modo não se trata de um valor nulo ou zero). Também a coluna *aecn*, mostra em todas as linhas sempre o mesmo valor, situação que pode conduzir à irrelevância da coluna alvo de análise.

Em seguida, são enunciadas, num primeiro momento, as anomalias verificadas nos dados desta tabela, a sua catalogação e o método para a sua resolução (tabela 6-18). As observações são resultantes das informações fornecidas pelo software *Datiris* (tabela 6-16 e 6-17) e da observação directa dos dados. Num momento posterior, é realizada uma análise para aferir da existência de linhas duplicadas na tabela. Uma avaliação cuidada sobre os resultados obtidos pelo software *Datiris*, sobre a vista 2, aponta para um valor na ordem dos 14% de linhas com a coluna *razao* com valores repetidos e a existência de cerca de dois terços de linhas com padrões distintos na coluna referida. Este último aspecto é importante, nomeadamente, na divulgação de relatórios e respostas às solicitações dos consumidores. Quanto à duplicação de valores é possível observar na vista 4, alguma proximidade semântica entre linhas da tabela, conforme se pode observar na tabela 6-19.

| Coluna | Descrição | Problema | Valor original | Método | Valor a obter |
|---|---|---|---|---|---|
| Aecn | Presença exclusiva do mesmo valor (cf. resultados de *data profiling)*. | 2 | 999999 | 11 | |
| Moeda | Moeda em Espanha é euro e não esp. | 4.2.2 | Esp<br>FRF<br>ITL | 1.3<br>8 | EUR<br>EUR<br>EUR |
| Moeda | Valores errados (empresa portuguesa negoceia em USD?) | 4.2.2 | MALHAS SEREIA,SA; PT; USD | 2.3<br>7 | MALHAS SEREIA, SA; PT; EUR |
| Nome_comercial | Valores omissos ou incompletos. | 4.1.2.2 | ZZ COMPRAS A DINHEIR<br>ZIMUTE-APRESTOS MARI | 2.1<br>2.2 | ZZ COMPRAS A DINHEIRO<br>ZIMUTE-AP. MARITIMOS |
| Nome_comercial | Inserção de símbolos.<br>Valores sem significado | 4.2.4<br>4.2.7 | XAVIER FERREIRA & FºS*PP*<br>ZELNOVA*MP*<br>DIAMOND BLUE TRADING*PL*<br>ELECTR. TAURUS "MP" (USD) | 1<br>2.3 | |
| Nome_comercial | Contém valores além do contexto da coluna | 4.2.4 | PINHEIRO ARROJA-TRANSF.<br>PNR LOURES - TRANSF.<br>PR GUIA-TRANSF.<br>HIJOS TIMOTEO RUIZ (TAO) | 1<br>11 | |
| Nome_comercial | Contém valores além do contexto da coluna. (valor repetido) indica o país ou a moeda | 2<br>4.2.4 | SERGIO TACCHINI (ESP)<br>DOHE(POR)<br>MEKA BLOCK(PT)<br>GILLETTE (ESPANA)<br>ETS J.RONDINAUD(FRF) | 1<br>11 | |
| Pais | Duplicação de valores para a mesma entidade | 4.2.3 | IN – GB<br>CH – CN | 8 | IN<br>CH |
| Pais | Pouco elucidativo (apenas 2 caracteres).<br>Como distinguir Escócia de Espanha? | 4.1.2.1 | ES<br>ES<br>PK; VG; MY | 11 | ESC<br>ESP |

**Tabela 6-18 – Anomalias dos valores dos dados na dimensão *fornecedor* (continua).**

| Coluna | Descrição | Problema | Valor original | Método | Valor a obter |
|---|---|---|---|---|---|
| Razão | Erros ortográficos | 4.2.1 | ADEGA COOP.MURCA,CRL<br>AGRIAL-AG.RIBATEJO ALENTRJO,SA<br>ABOLETA BARRANQUENHA (TAL | 2.3<br>8 | ADEGA COOP.MURÇA, CRL<br>AGRIAL-AG.RIBATEJO ALENTEJO, SA<br>A BOLETA BARRANQUENHA |
| Razão<br>Nome_comercial | Sem estandardização / normalização de valores e conceitos.<br>Regras de negócio | 4.1.2.3 | PINHEIRO ARROJA-TRANSF.<br>PNR LOURES - TRANSF.<br>PR GUIA-TRANSF.<br>LTD; LDA; LDA.; LTD. | 2.1<br>2.2 | PNR – ARROJA TRANSF.<br>PNR – LOURES<br>PNR – GUIA<br>LDA. |
| | Linhas duplicadas (pela aplicação *wizsame*). | 2 | ANGUS DUNDEE (4)<br>A.S.FASHION PRIVATE LIMITED (3)<br>ANIBAL RODRIGUES (2) | | |

**Tabela 6-18 – Anomalias dos valores dos dados na dimensão *fornecedor* (continuação).**

| MatchNum | RecordNum | FORNECEDOR_KEY | COD_FORNECEDOR | RAZAO | NOME_COMERCIAL |
|---|---|---|---|---|---|
| 2 | 11 | 44325.000000 | 11099 | 3A,SA | 3 A-CANDIA |
| 2 | 12 | 44722.000000 | 11184 | 3A,SA | 3 A SA CANDIA |
| 4 | 38 | 50223.000000 | 13647 | A COLMEIA MINHO,SA | A COLMEIA MINHO(F/L) |
| 4 | 39 | 50220.000000 | 14223 | A COLMEIA MINHO,SA | A COLMEIA MINHO*MP* |
| 4 | 40 | 51373.000000 | 00002 | A COLMEIA MINHO,SA | A COLMEIA MINHO,SA |
| 6 | 50 | 37492.000000 | 00011 | A METALURGICA-BAKEWARE PROD.,SA | A METALURGICA |
| 6 | 51 | 45578.000000 | 11400 | A METALURGICA-BAKEWARE PROD.,SA | A METALURGICA*MP* |
| 7 | 59 | 43836.000000 | 08165 | A PLUS MULTIMEDIA,LDA | A PLUS MULTIMEDIA |
| 7 | 60 | 44316.000000 | 11097 | A PLUS MULTIMEDIA,LDA | A PLUS MULTIMEDIA(INF) |
| 11 | 118 | 39974.000000 | 08016 | A.J.GONÇALVES,SA | A.J.GONÇALVES *MX* |
| 11 | 119 | 37514.000000 | 00033 | A.J.GONÇALVES,SA | A.J.GONÇALVES |
| 12 | 142 | 51105.000000 | 07632 | A.MONIZ-PROD.MAQ.EQUIP.,LDA | A.MONIZ (BARBOT) |
| 12 | 143 | 41197.000000 | 07484 | A.MONIZ-PROD.MAQ.EQUIP.,LDA | A.MONIZ |
| 12 | 144 | 42209.000000 | 12858 | A.MONIZ-PROD.MAQ.EQUIP.,LDA | A.MONIZ(BARBOT)*MP* |
| 13 | 151 | 49855.000000 | 05559 | A.PARODI,LDA | A.PARODI |
| 13 | 152 | 37529.000000 | 00048 | A.PARODI,LDA | A.PARODI,LD |
| 14 | 158 | 49521.000000 | 03955 | A.PIRES LOURENÇO & FILHOS,SA | A.PIRES LOURENCO & FºS |
| 14 | 159 | 37179.000000 | 12800 | A.PIRES LOURENÇO & FILHOS,SA | A.PIRES LOURENÇO*PP* |
| 15 | 169 | 37498.000000 | 00017 | A.S.DUARTE,LDA | A.S.DUARTE,LDA |
| 15 | 170 | 36638.000000 | 02821 | A.S.DUARTE,LDA | A.S.DUARTE |
| 15 | 171 | 42399.000000 | 06973 | A.S.DUARTE,LDA | A.S.DUARTE *MX* |
| 16 | 173 | 38159.000000 | 15870 | A.S.FASHION PRIVATE LIMITED | A.S.FASHION PRIVATE-PCOTT |
| 16 | 174 | 39570.000000 | 14647 | A.S.FASHION PRIVATE LIMITED | A.S.FASHION PRIVATE-PCOTT |

**Tabela 6-19 – Excerto de parte das linhas duplicadas da vista 4, da dimensão *fornecedor* (*wizsame)*.**

## Dimensão promoção

A análise dos dados da dimensão *promoção* foi realizada sobre três vistas materializadas (tabela 6-20). A primeira corresponde à totalidade da tabela e servirá de repositório base para a aplicação *Datiris*. A segunda vista é relativa às primeiras 999 linhas da tabela original e surge devido às limitações das aplicações *wizsame* e *wizrule*. A última vista considerada incide nas linhas que apresentam a coluna descrição igual a 'promoção' ou um valor idêntico. A definição desta vista deve-se ao elevado número de linhas que correspondem ao critério considerado e tem como objectivo determinar mais facilmente a existência de linhas duplicadas.

| Nº | Vista | Tamanho (em linhas) | Descrição |
|---|---|---|---|
| 1 | Promoção | 19 154 | Todas as linhas e colunas da tabela original. |
| 2 | Promoção_999 | 999 | Contém as primeiras 999 linhas da tabela original. |
| 3 | Promoção_pr | 550 | As linhas da tabela com a coluna *descricao* igual a 'promoção' ou similar. |

**Tabela 6-20 – Vistas definidas para avaliação sobre a dimensão *promoção*.**

Uma análise geral sobre a tabela revela a coluna *descrição* como aquela que apresenta o maior grau de probabilidade da ocorrência de defeitos nos valores dos dados. Esta observação resulta da heterogeneidade dos valores existentes na coluna e que consequentemente, deriva da ausência de regras de negócio e cumprimento de critérios na inserção destes valores na tabela. As informações obtidas pelo processo de *data profiling*, sobre a vista 1, corroboram esta análise. Conforme é observável na tabela 6-21, cerca de 27% das linhas apresentam valores iguais na coluna descrição e muitas outras linhas mostram valores similares ou sinónimos. Esta questão é merecedora de maior atenção na medida em que, praticamente, metade dos valores distintos (6970) exibe padrões de construção diversos. Esta tendência confirma a despreocupação na orientação sobre a implementação de padrões regulares na construção dos valores introduzidos. É, igualmente, possível observar a elevada taxa de valores ausentes nas colunas *evento*, *prom_type* e *old*, registando 47%, 57% e 100% respectivamente. Ainda no domínio da ausência de valores dos dados verifica-se a indistinção entre valores nulos e ausentes. Logo, a pertinência destas colunas no DM é determinada pelo seu efectivo preenchimento, caso contrário a sua presença indicia-se pouco relevante para os consumidores finais.

| Coluna | Distintos | | Nulos | Zeros | Brancos | Padrões | Observações |
|---|---|---|---|---|---|---|---|
| Descricao | 13 916 | 73% | | | | 6 970 | |
| Evento | | | 48% | | | | |
| Prom_type | | | 57% | | | | |
| Old | | | 100% | | | | Sem valores |

**Tabela 6-21 – Características da vista 1, da dimensão *promoção* (*datiris*).**

Em seguida, num primeiro momento, são enunciadas as imperfeições verificadas nos dados desta tabela, a sua catalogação e o método de resolução. As observações resultam de informações fornecidas pelo software *Datiris* (tabela 6-22) e da observação directa sobre os dados. Num momento posterior, é realizada uma análise para aferir da existência de linhas duplicadas na tabela. Os resultados obtidos pela ferramenta de *data profiling* na vista 1 identificam mais de um quarto de valores idênticos na coluna *descricao*. Esta situação merece uma análise mais cuidada em vista aferir sobre a possibilidade de duplicação de linhas de valores. A análise apenas sobre esta coluna não permite concluir sobre a possível duplicação de linhas, outras colunas necessitam ser confrontadas (*data_inicio* e *data_fim*). Para a elucidação desta questão recorreu-se à aplicação *wizsame*, com o intuito de aferir sobre a duplicação de linhas. A aplicação da ferramenta *wizsame* implicou a partição da tabela original em tranches de 1000 linhas, devido a limitações impostas pela versão do programa disponível e assim utilizou-se a vista 2. Tendo em vista determinar a duplicação de linhas, a aplicação foi parametrizada para as colunas: *descricao, data_inicio* e *data_fim*. Um pequeno excerto dos resultados obtidos, que indiciam a presença de linhas duplicadas, pode ser observado na tabela 6-23.

Dado que a coluna *descricao* apresenta um significativo número de linhas com conteúdo igual ou similar a 'promoção', optou-se por submeter a vista 3 à aplicação *wizsame*. Tendo em vista a obtenção de resultados mais objectivos, parametrizou-se a aplicação para considerar linhas duplicadas quando a coluna *descricao* apresente um conteúdo similar e as colunas *data_inicio* e *data_fim* sejam exactamente iguais (tabela 6-24). Obtiveram-se cerca de 80 ocorrências que satisfizeram os requisitos predefinidos e como tal indiciando a presença de valores duplicados. Uma possível justificação para a presença de linhas duplicadas deve-se à introdução de uma ocorrência por cada unidade funcional que efectue uma promoção. Pelo que é dado a observar, uma mesma promoção pode estar a ocorrer simultaneamente em unidades funcionais distintas. Esta realidade inviabiliza uma análise mais flexível e de maior grau de profundidade nos acessos realizados pelos consumidores dos dados. Por exemplo, como se determina a influência da linha 'Saldo montanha 50%', em cada unidade funcional ou artigo, se existem dezasseis linhas iguais?

A confirmação relativamente a esta suposição apenas pode ser dissipada pelo confronto com as ocorrências da tabela de factos. Todavia, por razões anteriormente invocadas esse estudo não é possível ser realizado, ficando meramente a recomendação para a sua avaliação.

| Coluna | Descrição | Problema | Valor original | Método | Valor a obter |
|---|---|---|---|---|---|
| Descrição | Erros ortográficos | 4.2.1 | promossões chapeus<br>promossoes mochilas<br>promocaõ relógios<br>Promoçõa Set Golfe | 2.3<br>8 | PROMOÇÃO CHAPÉUS<br>PROMOÇÃO MOCHILAS<br>PROMOÇÃO RELÓGIOS<br>PROMOÇÃO SET GOLFE |
| Descrição | Valores em maiúsculas e minúsculas (usar tipo de letra mono-espaçada para facilitar nas listagens). | 4.1.2.3 | SALDOS<br>Saldos<br>Saldo | 2.1 | SALDOS<br>SALDOS<br>SALDOS |
| Descrição | Valores omissos ou incompletos. | 4.1.2.2 | SALDOS – LUA<br>SALDOS<br>REGRESSO ÁS AULAS-LSS ILU | 2.1<br>9 | SALDOS – LUA<br>SALDOS – NEVE<br>REG. AULAS-LSS ILUMINAÇÃO |
| Descrição | Valores ambíguos | 4.1.2.1 | Sal<br>sporting 1; sporting 2; sporting<br>1; 1; 1<br>12.9; 12.9; 12.9 | 7 | Sal ou saldos? |
| Descrição | Valores sem significado<br>Inserção de símbolos | 4.2.7 | thjdj; vv; z<dv; hJJJJ<br>mario; ¿ 19,90; c; ç | 2.3 | |
| Descrição | Sem estandardização / normalização de valores e conceitos. | 4.1.2.3 | ADI-FEIRA PRAIA DCA LSS<br>ADIT FOLHETO G 18 LS<br>ADIT. ABERT. ANADIA<br>ADITAMENTO AZEITE LS<br>1490; 14,9; 14.90¿; 14.90  2; 14.9 | 2.1<br>2.2 | ADIT. FEIRA PRAIA DCA LSS<br>ADIT. FOLHETO G 18 LS<br>ADIT. ABERTURA ANADIA<br>ADIT. AZEITE LS |
| Descrição<br>Tipo_gestao<br>Variante<br>Prom_type | Tem espaços em branco antes e após o valor (cf. com os resultados obtidos por *data profiling*). | 4.1.2.3 | _DCBL_ADIT_P9_LS_2004____<br>1002___1990_____<br>PINHEIRO_-_6MESES_____ | 2.1 | DCBL ADIT P9 LS 2004<br>1002 1990<br>PINHEIRO – 6 MESES |

**Tabela 6-22 – Anomalias dos valores dos dados da vista 1 na dimensão *promoção*, (*Datiris*) (continua).**

| Coluna | Descrição | Problema | Valor original | Método | Valor a obter |
|---|---|---|---|---|---|
| Descrição | Contém valores além do contexto da coluna. (colocar uma coluna para unidade funcional ou tipo de loja, etc.). | 4.2.4 | Saldos 07/01/04 a 28/02/0<br>RELAMPAGO 7 LUA<br>RELAMPAGO 7 SOL<br>14.90 loja 179 Gondola pr | 1<br>11 | |
| Evento<br>Prom_type<br>Old | Presença massiva de valores nulos nas colunas (cf. resultados de *data profiling)*.<br>Aparente indistinção entre valores ausentes e nulos.<br>Sem valores por defeito. | 1.2 | Espaço ? | 9<br>3<br>11 | ? |
| Evento | Contém acrónimos, códigos e descrições.<br>Sem estandardização de conceitos e regras de negócio | 4.1.2.5 | MS0506<br>TESTBP<br>SALDOS<br>P60101<br>230103 | 2.1<br>2.2<br>8 | |
| | Linhas duplicadas | 2 | saldo 50% txt montanha (8)<br>12.9 (12)<br>c (9)<br>shopping (15) | | |
| Tipo_gestão | Contém apenas ‘MARKETING’ e ‘PONTUAL’. | 4.1.2.6 | MARKETING; PONTUAL | 11 | ‘M’ e ‘P’; 0 e 1 |

**Tabela 6-22 – Anomalias dos valores dos dados da vista 1 na dimensão *promoção*, (*Datiris*) (continuação).**

| MatchNum | RecordNum | PROMOCAO_KEY | CODIGO | DESCRICAO | DATA_INICIO | DATA_FIM | TIPO_GESTAO |
|---|---|---|---|---|---|---|---|
| 1 | 3 | 14720.000000 | 9401.000000 | TEXTIL LAR-94 | 01-01-1996 | 31-12-1996 | MARKETING |
| 1 | 26 | 14743.000000 | 9511.000000 | TEXTIL LAR | 01-01-1996 | 31-12-1996 | MARKETING |
| 1 | 45 | 14762.000000 | 9552.000000 | LAR | 01-01-1996 | 31-12-1996 | MARKETING |
| 2 | 4 | 14721.000000 | 9402.000000 | ANIVERSÁRIO | 01-01-1996 | 31-12-1996 | MARKETING |
| 2 | 39 | 14756.000000 | 951060.000000 | ANIVERSARIO | 01-01-1996 | 31-12-1996 | MARKETING |
| 2 | 96 | 14813.000000 | 95999.000000 | ANIVERSÁRIO - AUTOMÓVEIS | 01-01-1996 | 31-12-1996 | MARKETING |
| 2 | 137 | 14854.000000 | 961504.000000 | ANIVERSÁRIO | 01-01-1996 | 31-12-1996 | MARKETING |
| 3 | 5 | 14722.000000 | 9403.000000 | BÉBÉ | 01-01-1996 | 31-12-1996 | MARKETING |
| 3 | 103 | 14820.000000 | 960301.000000 | Bébé | 01-01-1996 | 31-12-1996 | MARKETING |
| 4 | 6 | 14723.000000 | 9404.000000 | PÁSCOA | 01-01-1996 | 31-12-1996 | MARKETING |
| 4 | 43 | 14760.000000 | 951070.000000 | PASCOA | 01-01-1996 | 31-12-1996 | MARKETING |
| 5 | 11 | 14728.000000 | 9409.000000 | FÉRIAS | 01-01-1996 | 31-12-1996 | MARKETING |
| 5 | 115 | 14832.000000 | 960901.000000 | Férias (Campo) | 01-01-1996 | 31-12-1996 | MARKETING |
| 5 | 119 | 14836.000000 | 961003.000000 | Férias (Praia) | 01-01-1996 | 31-12-1996 | MARKETING |
| 6 | 12 | 14729.000000 | 9410.000000 | PREÇOS NOTÍCIAS | 01-01-1996 | 31-12-1996 | MARKETING |
| 6 | 13 | 14730.000000 | 9411.000000 | PREÇOS | 01-01-1996 | 31-12-1996 | MARKETING |
| 7 | 14 | 14731.000000 | 9412.000000 | REGRESSO ÀS AULAS | 01-01-1996 | 31-12-1996 | MARKETING |
| 7 | 72 | 14789.000000 | 95121.000000 | REGRESSO AULAS | 01-01-1996 | 31-12-1996 | MARKETING |
| 8 | 15 | 14732.000000 | 9413.000000 | OUTONO/INVERNO | 01-01-1996 | 31-12-1996 | MARKETING |
| 8 | 77 | 14794.000000 | 95132.000000 | OUTONO | 01-01-1996 | 31-12-1996 | MARKETING |
| 9 | 17 | 14734.000000 | 9415.000000 | BRINQUEDOS | 01-01-1996 | 31-12-1996 | MARKETING |
| 9 | 88 | 14805.000000 | 95153.000000 | BRINQUEDOS | 01-01-1996 | 31-12-1996 | MARKETING |
| 9 | 130 | 14847.000000 | 961701.000000 | Brinquedos | 01-01-1996 | 31-12-1996 | MARKETING |

**Tabela 6-23 – Excerto de parte das linhas duplicadas na vista 2 da dimensão *promoção* (*Wizsame*).**

| MatchNum | RecordNum | PROMOCAO_KEY | CODIGO | DESCRICAO | DATA_INICIO | DATA_FIM | TIPO_GESTAO |
|---|---|---|---|---|---|---|---|
| 3 | 47 | 20796.000000 | 2217912.000000 | PROM BONES | 20-04-2003 | 31-05-2003 | MARKETING |
| 3 | 48 | 20797.000000 | 2217913.000000 | PROM BONES 1 | 20-04-2003 | 31-05-2003 | MARKETING |
| 4 | 66 | 20962.000000 | 2219233.000000 | Promoção Calçado 10/05/03 | 11-05-2003 | 30-06-2003 | MARKETING |
| 4 | 67 | 20964.000000 | 2219273.000000 | Promoçao Calçado 10/05/03 | 11-05-2003 | 30-06-2003 | MARKETING |
| 5 | 73 | 21044.000000 | 2219653.000000 | promoção golfe | 17-05-2003 | 30-06-2003 | MARKETING |
| 5 | 74 | 21045.000000 | 2219657.000000 | promoção golfe | 17-05-2003 | 30-06-2003 | MARKETING |
| 7 | 77 | 21050.000000 | 2219693.000000 | Promoção Sporting | 17-05-2003 | 31-07-2003 | MARKETING |
| 7 | 78 | 21051.000000 | 2219694.000000 | Promoção Sporting | 17-05-2003 | 31-07-2003 | MARKETING |
| 7 | 79 | 21052.000000 | 2219695.000000 | Promoção Sporting | 17-05-2003 | 31-07-2003 | MARKETING |
| 18 | 166 | 21827.000000 | 2224964.000000 | promoção nike | 16-08-2003 | 31-08-2003 | MARKETING |
| 18 | 167 | 21828.000000 | 2224965.000000 | PROMOÇÃO NIKE | 16-08-2003 | 31-08-2003 | MARKETING |
| 18 | 168 | 21829.000000 | 2224966.000000 | PROMOÇÃO NIKE | 16-08-2003 | 31-08-2003 | MARKETING |
| 19 | 176 | 22379.000000 | 2228213.000000 | promo meias com embalagen | 17-10-2003 | 30-11-2003 | MARKETING |
| 19 | 177 | 22380.000000 | 2228214.000000 | promo meias com embalagen | 17-10-2003 | 30-11-2003 | MARKETING |
| 19 | 178 | 22381.000000 | 2228215.000000 | promo meias com embalagen | 17-10-2003 | 30-11-2003 | MARKETING |
| 19 | 179 | 22382.000000 | 2228223.000000 | promo meias com embalagen | 17-10-2003 | 30-11-2003 | MARKETING |
| 19 | 180 | 22383.000000 | 2228224.000000 | promo meias com embalagen | 17-10-2003 | 30-11-2003 | MARKETING |
| 19 | 181 | 22384.000000 | 2228233.000000 | promo meias com embalagen | 17-10-2003 | 30-11-2003 | MARKETING |
| 27 | 241 | 23316.000000 | 2234893.000000 | prom combate | 25-01-2004 | 29-02-2004 | MARKETING |
| 27 | 242 | 23317.000000 | 2234894.000000 | prom combate | 25-01-2004 | 29-02-2004 | MARKETING |
| 27 | 243 | 23318.000000 | 2234895.000000 | prom combate | 25-01-2004 | 29-02-2004 | MARKETING |
| 27 | 244 | 23319.000000 | 2234896.000000 | prom combate | 25-01-2004 | 29-02-2004 | MARKETING |
| 27 | 246 | 23321.000000 | 2234898.000000 | prom combate | 25-01-2004 | 29-02-2004 | MARKETING |

**Tabela 6-24 – Excerto de parte das linhas duplicadas na vista 3 da dimensão *promoção* (*Wizsame*).**

## 6.5 Algumas recomendações

O objectivo da apresentação de um conjunto de recomendações de melhoria da qualidade dos dados do SDW, consiste em estabelecer uma plataforma capaz de permitir a gestão dos dados existentes e perspectivar a evolução da organização em termos de maturidade e domínio dos dados organizacionais. Neste sentido, as iniciativas de recomendação que visam debelar os defeitos existentes nos dados do SDW podem ser enquadradas desde iniciativas de âmbito estratégico até outras de âmbito estritamente operacional.

### 6.5.1 Recomendações de cariz estratégico

As recomendações estratégicas são iniciativas de reestruturação dos processos e actividades organizacionais visando a melhoria dos dados existentes no SDW. Por isso, correspondem a horizontes temporais alargados e apresentam-se transversais a toda a organização. O reconhecimento dos dados como um assunto organizacional e importante na sua condução estratégica constitui uma condição essencial tendo em vista o tratamento efectivo desta problemática. A enraização do conceito de informação como um produto acabado composto por um conjunto de características, executado por processos de transformação específicos e servindo os desejos dos consumidores finais, bem como a adopção duma metodologia que promova a qualidade dos dados (TDQM), mostra-se como um ponto de partida inicial. O entendimento dos dados como recurso estratégico determina a definição de estratégias para os dados. Assim, iniciativas de promoção da melhoria contínua dos dados, assentes em meios de antecipação e prevenção da ocorrência de problemas a montante, simplificam as operações a jusante. Este tipo de medidas pressupõe o cumprimento de um conjunto de pressupostos, como sejam: a mudança cultural (a promoção de políticas de incentivos e formação aos diversos colaboradores), o desenvolvimento de um sistema de gestão dos dados alargado a toda a organização, a reestruturação ou substituição dos sistemas informáticos existentes (meios de captação dos dados mais precisos e intuitivos). Subjacente a estas questões, depreende-se a existência de uma área funcional própria que exerça as suas actividades em torno dos dados organizacionais, dirigida por responsáveis capazes de cobrir os diferentes aspectos relacionados com os dados (administradores dos dados, responsáveis pelos metadados, guardiães dos dados) e capaz de implementar e coordenar programas por si emanados. Na prática, promover a engenharia dos dados organizacionais.

A perspectiva estratégica, relativamente aos dados, consubstancia-se na antevisão da organização num futuro mais ou menos longínquo e no enquadramento dos dados como a fonte de proveniência de vantagens concorrenciais. O alinhamento da exploração dos dados à estratégia organi-

zacional, permite estabelecer o SDW como a plataforma tecnologicamente dotada para o processamento analítico dos dados, em especial, quando associada a meios e técnicas, como por exemplo a mineração dos dados. Estes meios, ainda inexistentes na organização, permitem, quando devidamente explorados, a criação de conhecimento e sabedoria e deste modo, catapultar a organização para uma dimensão mais competitiva (aproveitar oportunidades e corrigir fraquezas). Porém, os dados actualmente disponibilizados não se encontram em condições deste propósito. Em suma, a orientação geral das recomendações consiste em promover uma gestão eficaz e eficiente dos dados residentes no SDW e deste modo, dotar esta plataforma para cenários mais dinâmicos e exigentes no futuro.

### 6.5.2 Recomendações de cariz operacional

Do ponto de vista operacional, as iniciativas a tomar desenvolvem-se em dois tipos de actividades: as de âmbito externo aos SDW e as de âmbito restrito ao SDW. As primeiras centram-se em medidas que promovem a captura correcta dos valores nas fontes, como sejam, entre outras, o desenvolvimento de formulários mais intuitivos e restritivos para a entrada correcta dos dados, a formação dos colaboradores e a adopção de meios automáticos ou semi-automáticos de captura dos dados. Ainda sobre esta classe de medidas, importa o levantamento das características dos dados a respeitar, através de uma engenharia de requisitos adequada (e.g. método QFD). Esta iniciativa permite perspectivar os critérios de qualidade que os dados devem corresponder (qualidade de projecto e qualidade de conformidade). Enquanto, as segundas actividades referem-se ao processo de ETL, ao modo de disponibilização dos dados aos consumidores finais e à monitorização do sistema. O facto da organização não possuir uma verdadeira ARD, associada à inexistência de um repositório do DW implica necessariamente que sejam as primeiras questões a resolver.

A divulgação de dados a um leque alargado de utilizadores, o peso histórico do sistema e o enorme volume de dados registado, condicionam a manutenção de DMs flexíveis e orientados para os agentes de decisão. Por esse motivo, parece-nos que, o facto da concepção do SDW existente assentar numa arquitectura em *bus* pura [Kimball e tal, 1998], não responde cabalmente às exigências organizacionais. A implementação de um repositório de DW precedente à difusão dos dados pelos DMs, congregando todos os dados passíveis de interrogações, permitiria uma maior homogeneização e agregação dos dados pelos diversos DMs. Deste modo, reduzir-se-á a propensão de dados dormentes e disparatados nos repositórios acedidos pelos agentes de decisão.

No que respeita à ARD, importa referir que se trata do local apropriado para impor níveis de qualidade aos dados a integrar no DW porque corrige e acrescenta valor aos dados existentes e contri-

bui decisivamente na disponibilização dos dados mais relevantes no mais curto espaço de tempo, de forma acessível e facilmente interpretável. A utilização da ARD permite obter no DW e nos DMs uma versão consistente dos dados do SO. Os meios e recursos consumidos neste local correspondem geralmente a cerca de dois terços do orçamento previsto para a implementação e manutenção de um SDW, situação certamente oposta à vivida pela organização em estudo. As rotinas actualmente existentes não se encontram pressionadas por imposições temporais não exequíveis durante a janela de oportunidade existente. Acresce que, hoje em dia, algumas rotinas visando a melhoria da qualidade dos dados de determinadas tabelas já são executadas (e.g. remoção de espaços em branco na dimensão *fornecedor*), por isso, alastrar estas rotinas às restantes tabelas parece uma trivialidade. Porém, conforme foi dado a observar, um conjunto de outras iniciativas merece especial atenção e que podem ser segmentadas em três fases: os dados a captar do SO (requisitos dos utilizadores, análise dos dados, etc.), as operações de transformação e limpeza dos dados recolhidos (adequar os instrumentos para remoção de anomalias) e o modo como os dados tratados são carregados em vistas de acesso pelos consumidores finais.

## Primeira fase – dados a extrair do SO

A recolha dos dados a extrair das fontes é fortemente condicionada pela existência dos dados nas fontes e pela engenharia de requisitos realizada junto dos consumidores. Assim, o primeiro passo consiste em confrontar os requisitos dos consumidores com os dados possíveis de obter no SO. A análise das características inerentes aos dados pode ser realizada utilizando para o efeito uma ferramenta de *data profiling*. Da análise às fontes deve constar, igualmente, a determinação dos seus perfis, em termos de actualização e disponibilização para extracção dos dados.

## Segunda fase – operações de transformação e limpeza dos dados

As operações de transformação e limpeza dos dados devem ocorrer sobre os dados em trânsito para a ARD e compreendem um conjunto de operações em vista a remoção de anomalias e consequentemente, a obtenção de um repositório de dados mais fielmente representativo do mundo real. As operações são executadas sequencialmente e consistem na reparação das imperfeições diagnosticadas nos dados do SDW, conforma mostra a figura 6-6.

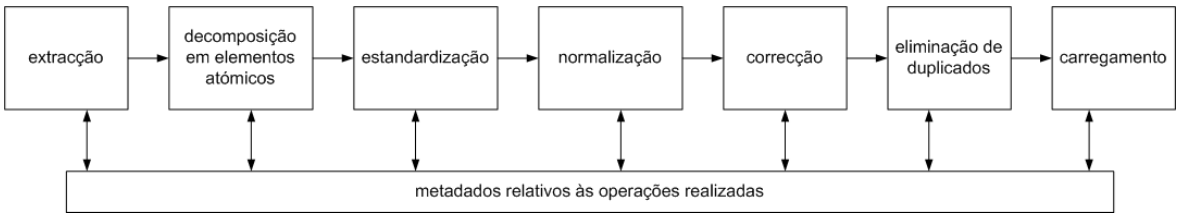



**Figura 6-6 – Etapas de transformação e limpeza dos dados.**

**Decomposição dos dados em elementos atómicos**

A decomposição dos valores em elementos atómicos consiste em captar e redireccionar para colunas específicas os valores das colunas, de formato livre ou sobrecarregadas de valores no seu conteúdo, provenientes do SO (tabela 6-25).

| Tabela | Dados originais | Decomposição por elementos | |
|---|---|---|---|
| Fornecedor | ELECTR. TAURUS "MP" (USD) | Nome_comercial: | ELECTR. TAURUS |
| | | Moeda: | USD |
| | | *Identificador de coluna*: | MP |
| Fornecedor | SERGIO TACCHINI (ESP) | Nome_comercial: | SERGIO TACCHINI |
| | | Pais: | ESP |
| Promoção | RELAMPAGO 7 LUA | Descrição: | RELAMPAGO 7 |
| | | *Identificador de coluna*: | LUA |
| Promoção | RELAMPAGO 7 SL | Descrição: | RELAMPAGO 7 |
| | | *Identificador de coluna*: | SL |

**Tabela 6-25 – Exemplos da operação de decomposição dos dados em elementos atómicos.**

Após a decomposição em valores atómicos proceder-se à execução de outras operações, como sejam: a correcção (e.g. SL em SOL), a validação (e.g. a moeda é PTE), o preenchimento de valores ausentes, a normalização e a estandardização de valores dos dados.

**Estandardização dos dados**

No que respeita à estandardização dos dados devem ser definidos os valores que melhor representam, de forma imediata e ausente de dúvidas, um conceito ou entidade (tabela 6-26).

| Dados originais | Standard escolhido | Dados estandardizados |
|---|---|---|
| Promo. | | Promoção |
| Promoção | Promoção | Promoção |
| Promoções | | Promoção |
| LTD. | | LDA. |
| LDA. | LDA. | LDA. |
| Lda | | LDA. |

**Tabela 6-26 – Exemplos de estandardização dos dados.**

**Normalização dos dados**

Na normalização de valores devem ser tidas em conta as regras de negócio e os critérios a respeitar (e.g. 3 caracteres – localidade) (tabela 6-27).

| Dados originais | Dados normalizados |
|---|---|
| PR VISEU | PNR – VISEU |
| PNR-MARINHA GRANDE | PNR – MARINHA GRANDE |
| PNR LOURES | PNR – LOURES |
| PINHEIRO ESPOSENDE | PNR – ESPOSENDE |
| MC-CACEM | MC-CACEM |
| MES-VISE | MC-VISEU |
| MESA CAB (Leiria) | MC-LEIRIA |
| MS-BRAGA | MC-BRAGA |
| 467-VASC | MC-VASCO |

**Tabela 6-27 – Exemplos de normalização dos dados.**

### Correcção dos dados

Após a normalização e estandardização, segue-se a correcção e validação dos valores dos dados que compreende a rectificação de valores que violam o domínio de valores definido, as inconsistências de valores, a imposição das regras impostas pelo negócio, a remoção dos espaços em branco nos dados e o preenchimento de valores ausentes ou incompletos (tabela 6-28).

| Dados originais | | Dados correctos | |
|---|---|---|---|
| Razao: | JIN TAI IMP.EXP.ENTERPRISE GROUP CO | Razao: | JIN TAI IMP.EXP.ENTERPRISE GROUP CO |
| Nome: | JIN TAI | Nome: | JIN TAI |
| Pais: | CH | Pais: | CHI |
| Razao: | JIN TAI AND EXP.ENTREPRISE GROUP.CO | Razao: | JIN TAI AND EXP.ENTREPRISE GROUP.CO |
| Nome: | JIN TAI IMP.EXP. | Nome: | JIN TAI IMP.EXP. |
| Pais: | CN | Pais: | CHI |

**Tabela 6-28 – Exemplos de dados corrigidos.**

Uma outra classe de operações deve ocupar-se da transformação entre os tipos de dados adoptados ou entre a uniformização do domínio de valores possível. No caso em estudo verifica-se a adopção de terminologia portuguesa e inglesa na atribuição de identificadores de coluna (e.g. *prom_type*, *old*) e no domínio de valores possíveis para a coluna ('Y'). Por último, devem ser aplicadas rotinas para verificar a existência de linhas de dados duplicadas no repositório dos dados.

## Terceira fase – dados no DM

Os dados a carregar no DM de vendas devem ter em linha de conta as necessidades e desejos dos consumidores finais. Assim, devem constar os dados objectivamente relevantes para o processamento analítico dos dados, isto é, a materialização de vistas deve realizar-se de acordo com as necessidades ou perfis dos clientes. A opção em construir subconjuntos das linhas das dimen-

sões revela-se uma iniciativa a implementar e corroborada pelas tendências mais recentes das investigações. A manutenção no DM de dados dormentes, nunca ou raramente acedidos implica, necessariamente, a maior probabilidade de inconsistências, a perda de desempenho e o desperdício de recursos consumidos (humanos, materiais, energéticos, temporais, etc.). Neste sentido, deve ser ponderada a agregação de valores, a selecção das linhas e a projecção das colunas úteis nas respostas aos consumidores. De qualquer modo, continuarão a coexistir e igualmente acessíveis os dados não presentes no DM, mas numa camada precedente a esta.

O sucesso das diferentes fases abordadas apenas é verificável com a manutenção de metadados referentes às operações realizadas, mas igualmente, sobre as diversas camadas do SDW, os processos de circulação dos dados, histórico e taxas de utilização destes, as regras de negócio e os mapeamentos dos dados, desde o SO até ao DW. Os metadados assumem particular importância na qualidade dos dados visto que ajudam na identificação das quebras de linhagem nos dados. Conforme referido anteriormente, os metadados assumem-se como o modo de garantir a qualidade após a entrega dos dados. Portanto, os metadados são uma componente determinante na gestão da qualidade abrangente a todo o SDW e contribuem para o sucesso deste.

Algumas ferramentas comerciais encontram-se disponíveis no mercado, como sejam o *dfPowerStudio*, disponibilizado pela *Dataflux* [9] e o *Teradata Warehouse* e *Teradata Warehouse Miner*, da *NCR* [10]. Ambas procuram assegurar as tarefas de *back-end* do SDW e compreendem a generalidade das actividades inerentes ao processo de ETL. O pacote *dfPowerStudio* estabelece a gestão dos dados em torno da sequenciação de cinco grandes actividades: o *profiling*, a qualidade, a integração, o enriquecimento e a monitorização dos dados. As aplicações da *NCR* disponibilizam, igualmente, um conjunto de componentes: motor de regras, *profiling*, auditoria e operações transformação e limpeza.

A monitorização do sistema deve igualmente assentar na averiguação sobre o cumprimento dos critérios e desejos assumidos pelos consumidores dos dados. Devem ser realizados questionários que possibilitem aferir sobre o grau de cumprimento do sistema e realcem os aspectos merecedores de ajustamento. Os resultados obtidos por questionários devem ser cruzados com medidas objectivas dos dados, estrategicamente colocadas no sistema (e.g. *data profiling* sobre o DW), e consequentemente proceder ao alinhamento do sistema com a estratégia de negócio. A monitorização compreende a recolha de informações relativas às taxas de consulta aos dados em vista ponderar sobre:

– A determinação da probabilidade de acesso aos dados.

- A determinação dos dados que têm ou não sido acedidos.
- A criação de um perfil de acesso baseado na actividade passada.
- A atribuição de uma probabilidade de acesso baseada no perfil que é criado.
- A elaboração de uma lista que sistematize as prioridades na ausência de anormalidades nas colunas, determinando-se assim, as colunas críticos.

## 6.6 Comentários finais

O estudo deste caso abordou a problemática da qualidade dos dados num contexto real. Assim, as observações realizadas na amostra disponibilizada, sobre os dados do SDW alvo de estudo, enquadram a organização num panorama em que pouca atenção é prestada à qualidade dos seus repositórios de dados. Esta realidade justifica a atribuição do segundo nível de maturidade, dos cinco níveis existentes, em relação ao tratamento das questões dos dados. Os motivos para a atribuição deste nível deveram-se à escassez de iniciativas que prevejam a detecção, reparação e melhoria dos dados existentes. A inexistência de um programa transversal à organização que estabeleça este tema como um assunto organizacional, orientado por responsáveis que definem padrões e regras de conduta no manuseamento dos dados, agrava ainda mais o cenário existente. É igualmente observável a falta de metadados descritivos dos processos, das camadas da arquitectura e dos dados envolvidos, bem como, a ausência de ferramentas que tratem dos defeitos dos dados de modo sistemático. A auscultação dos dados constantes nos repositórios permite catalogar as diversas irregularidades que povoam o SDW e consequentemente, definir os melhores métodos de correcção dessas ocorrências. Estas acções geralmente consideradas na arquitectura dos SDW na ARD, não são convenientemente implementadas. Deste modo, os dados que povoam os DMs encontram-se feridos nas suas mais elementares características.

O trabalho realizado foi estruturado em quatro fases distintas. A primeira correspondeu à esquematização dos processos envolvidos desde o SO até ao DM alvo de estudo. A segunda baseou-se em observações gerais sobre os dados, entrevistas e inquéritos de modo a identificar o nível de maturidade da organização em relação ao tratamento dos dados. No terceiro momento foram realizadas análises exaustivas aos dados, determinando as suas características e estatísticas reveladoras dos defeitos existentes nos dados. Estas informações recolhidas confirmaram a atribuição do nível de maturidade proposto. Tendo em vista a classificação da natureza dos defeitos nos dados foi adoptada uma taxionomia de identificação de anomalias nos dados no DM do caso em estudo, bem como o modo de resolução dessas anomalias. Por último, procedeu-se a um conjunto de recomendações a implementar visando a melhoria da qualidade dos dados existentes no SDW.

Estas recomendações efectuaram-se tanto num plano estratégico e abrangente à totalidade da organização, como num plano operacional orientado para as questões centrais de um ambiente de DW e assentes numa metodologia que promova a melhoria contínua da qualidade dos dados.

Relativamente, aos trabalhos futuros possíveis de realizar destacam-se os custos provocados pela fraca qualidade dos dados; a análise de ROI a obter pela implementação de iniciativas de melhoria da qualidade dos dados e a definição de critérios a satisfazer junto dos agentes de decisão. Em suma, os resultados obtidos revelam uma qualidade dos dados abaixo dos padrões mínimos de exigência para uma organização desta grandeza, tendo em conta o sistema de dados utilizado neste caso de estudo. Neste contexto, a organização não se encontra preparada para enfrentar cenários mais exigentes em termos de qualidade dos dados, como seja, a introdução de técnicas e meios de mineração de dados, no modo de gestão do seu negócio. Assim e prevendo-se a acentuada necessidade de uso e utilidade dos dados organizacionais, na condução do seu negócio, procedemos à enunciação de algumas medidas ou estratégias de recomendação.

# Capítulo 7

# Conclusões e Trabalho Futuro

Ao longo da dissertação, tentámos demonstrar algumas deformidades verificáveis na qualidade dos dados no seio organizacional e, correlativamente, em ambientes de DW e que são geradoras de consequências nefastas para a organização e delimitadoras no sucesso dos SDWs.

Assim, este trabalho surge como corolário, numa primeira fase, da análise e reflexão da problemática da qualidade dos dados no campo organizacional, em geral, e em SDWs, em particular. Depois, da apreensão duma reunião de conceitos e terminologias básicas, bem como um conjunto de técnicas, metodologias, modelos e estratégias com o intuito de promover a melhoria da qualidade dos dados existente. Numa etapa posterior, após a consciencialização da natureza do trabalho e dos objectivos propostos, na transposição desses conceitos e assuntos apreendidos para a realidade concreta dos SDWs, através duma acção concertada dos objectos alvo de estudo e que se consubstanciou, ao que designamos, por uma plataforma de um sistema de gestão da qualidade dos dados em SDWs. Constatámos ser este um aliciante desafio a enfrentar, tanto pelo carácter actual e emergente embutido na própria natureza do trabalho, como também pelo facto de podermos encarar e perspectivar o sucesso dos SDWs em termos da qualidade das informações e dos indicadores divulgados, quer ainda pela possibilidade da investigação de um assunto ainda pouco prioritário nos trabalhos científicos e nos meios profissionais, em especial, oriundos do território nacional. Pretendemos com esta dissertação cooperar, nos meios académicos e profissionais, tendo em vista um conhecimento mais profundo da realidade relacionada com a qualidade dos dados em ambientes de DW e consequentemente na organização.

Este processo de investigação desencadeou uma vontade de nos inteirarmos de estudos e investigações entretanto desenvolvidas sobre o tema da qualidade dos dados. Constatámos ao longo da execução desta dissertação que, apesar da reconhecida importância dos dados nas organizações e nos DWs que servem de suporte aos agentes de decisão, estes são considerados por uma enorme franja de estudiosos deste fenómeno, como não cumprindo cabalmente o seu papel. Estes sistemas necessitam de uma profunda inovação nos processos de gestão das actividades de tratamento dos dados e no deslocamento da atenção por parte dos responsáveis organizacionais para uma visão do assunto num âmbito transversal à própria organização [English, 2003b] [Eckerson, 2002] [Shankaranarayan et al., 2000].

O tema escolhido para este trabalho "A Gestão da Qualidade dos Dados em Ambientes de DW na prossecução da Excelência da Informação" insere-se, certamente, num vasto universo de factos que atravessam toda a realidade dos SDWs e da própria organização. Se inicialmente, a qualidade dos dados dos SDWs não era vista como a fonte de deficiências operacionais, da perda de desempenho e do insucesso destes sistemas, estas eram justificadas sobretudo por deficiências de planeamento, ineficientes processos de refrescamento dos dados, desadequadas estruturas técnicas e tecnológicas, falhas graves na engenharia de requisitos, entre outros motivos. Posteriormente, pela insistência continuada dos casos de insucesso, mesmo após a resolução dos problemas anteriormente focados, constatou-se igualmente que a própria qualidade dos dados é originadora de falhas no cumprimento dos objectivos centrais destes sistemas e que se centram na disponibilização de mais e melhores informações aos agentes de decisão, permitindo uma maior elasticidade no acesso a mais dados, garantindo uma maior confiança sobre a informação disponibilizada e maior celeridade no processo de tomada de decisão. Esta realidade é demonstrada nos estudos [Watson et al., 2001] [Kenyon et al., 2000, 2004].

Paradoxalmente e apesar destes objectivos serem capazes de concederem os maiores benefícios às organizações são, muitas vezes, assumidos como os principais motivos de insucesso dos SDWs. É igualmente possível perspectivá-los numa óptica de qualidade dos dados e das dimensões associadas. As dimensões mais interessantes atendendo aos objectivos considerados e às características específicas dos SDWs correspondem ao cumprimento do grau de frescura dos dados fornecidos aos decisores, à acessibilidade e facilidade de interpretação dos dados divulgados, à utilidade e completude dos dados e culminando na exactidão e relevância das informações recebidas capazes de influenciar as decisões tomadas. O balanceamento equilibrado dos critérios definidos para as dimensões a satisfazer mostra-se como o modo adequado na gestão dos dados em ambientes de DW e dá provimento ao próprio conceito de qualidade dos dados. Assim, pretendeu-se com este trabalho salientar algumas incongruências exercidas no tratamento das ques-

tões relativas aos dados e analisá-las como estimuladoras de imperfeições nos dados pelas diversas camadas dos SDWs. Para enfrentar este desafio, estruturamos o estudo em torno de três momentos:

No primeiro momento, apresentámos a problemática da qualidade dos dados em redor das organizações e contextualizámos o assunto aos SDWs em concreto, para tal dispusemos de um conjunto de conceitos teóricos e princípios metodológicos, modelos e propostas de investigação que sustentam o tema desta dissertação, focando desde logo algumas limitações dos processos tradicionais de implementação dos SDWs, em especial, no que concerne à garantia da elevada qualidade dos dados nestes sistemas, bem como, as causas e a classificação dos defeitos verificados nos dados ao longo das diferentes camadas da arquitectura dos SDWs e métodos de tratamento desses defeitos. Este momento abarcou os primeiros três capítulos e as partes iniciais do quarto e quinto capítulos.

No segundo momento, procurámos descrever uma plataforma de um sistema de gestão da qualidade dos dados em ambientes de DW, baseada nos argumentos teóricos do momento precedente e que servem de base para o enriquecimento dos dados constantes no sistema (desde as fontes até à sua disponibilização aos agentes de decisão) de maneira a que este se apresente como suporte credível no processo de tomada de decisão. A plataforma assenta numa estratégia preventiva de antecipação da ocorrência de defeitos, através da compreensão dos problemas a montante, em detrimento da simples inspecção e reparação das anomalias verificadas e paralelamente na inserção de sistemas de medida espalhados pelos diversos patamares da arquitectura dos SDWs e que são capazes de recolherem informações sobre o estado da qualidade dos dados. Este momento congrega as segundas partes do quarto e quinto capítulos.

Nestes dois momentos não só estão apresentados os assuntos considerados como os mais relevantes, como também reflexões após a análise dos assuntos abordados. Por último, no que respeita ao terceiro momento e que corresponde ao sexto capítulo, foi apresentado um estudo de caso concreto sobre o DM de vendas de uma organização nacional real, que comprova a pouca atenção a que os dados se encontram sujeitos e que se sintetiza na atribuição do segundo nível da escala dos cinco níveis de maturidade dos dados referidos em [Adelman et al., 2005]. A abordagem sobre este estudo de caso assentou numa primeira fase na análise e diagnóstico dos processos que envolvem os dados e na classificação das anomalias verificadas, através da recolha de indicadores sobre as características dos dados. Por fim, estabeleceram-se um conjunto de algumas recomendações, assentes na plataforma por nós delineada, e que visou a melhoria da qualidade dos dados existente.

A principal contribuição desta dissertação consistiu numa melhor compreensão do fenómeno associado à problemática da qualidade dos dados em SDWs, nomeadamente, dos efeitos provocados pelas irregularidades dos dados, dos motivos que levam à presença dessas irregularidades no sistema e das estratégias existentes para minimizar essas irregularidades. As razões da presença de dados anómalos no sistema ou desadequados aos interesses dos decisores devem-se à captação de dados provenientes do SO de pouca qualidade, associadas a fontes heterogéneas, dispersas e antigas, a ineficazes políticas de tratamento desses dados e à presença de dados dormentes no DW. Os efeitos provocados pela presença de dados defeituosos incluem o desaproveitamento de oportunidades de negócio, a desconfiança e descrédito dos agentes de decisão sobre os dados divulgados e o aumento dos custos associados à necessidade de refazer o trabalho (custos da má qualidade) e outros custos consequentes da tomada de decisões desacertadas. De modo a minimizar os efeitos provocados, as organizações devem assumir os dados como um assunto organizacional, preparar antecipadamente a ocorrência de anomalias nos dados, promover uma gestão autónoma dos dados e elaborar uma estratégia para os dados assente numa metodologia que promova a melhoria contínua dos dados. A metodologia adoptada deve prever além da melhoria contínua dos dados, a adopção duma perspectiva sobre os dados como se de um produto convencional se tratasse, composto por características predefinidas e negociadas com os agentes de decisão (qualidade de conformidade) e processos de manuseamento capazes de fazer cumprir os critérios especificados (qualidade de conformidade).

Esta dissertação apresenta algumas limitações e deixa alguns assuntos relacionados com a problemática da qualidade dos dados em SDWs em aberto, que poderão ser resolvidos em trabalhos futuros. Seria interessante assim, como trabalhos futuros a investigação de temas como sejam:

- A determinação dos custos e benefícios originários da qualidade dos dados, em especial, no que respeita àqueles de maior dificuldade de obtenção (ocultos). Este conhecimento originário da boa ou má qualidade repercute-se no reconhecimento dos dados como importante recurso estratégico, nomeadamente pelo incentivo ao envolvimento por parte dos responsáveis organizacionais neste assunto.
- O desenvolvimento de mecanismos que prevejam a automatização ou processos semi-automáticos de entrada dos dados no SO e desse modo contribuam para a diminuição do principal factor de origem das deficiências nos dados.
- A identificação de alguns erros nos dados comuns de ocorrer em SDWs e que não se encontram catalogados pelas taxionomias de deformidades existentes. Nomeadamente, a presença de dados irrelevantes e dormentes no sistema.

- A captação automática de indicadores subjectivos sobre os níveis de qualidade dos dados divulgados aos agentes de decisão e posterior confronto com medidas objectivas no sentido de melhoria da qualidade dos dados e dos processos envolvidos.
- A realização de mais e melhores estudos sobre o tema e que simultaneamente realcem a problemática da qualidade dos dados em ambientes de DW.
- A promoção de uma área nos meios académicos e profissionais vocacionada para a engenharia dos dados.

A gestão dos dados deve ser consequência do reconhecimento da sua importância, por parte dos responsáveis organizacionais, através duma emancipação da sua administração autónoma. A divulgação de dados adequados aos agentes de decisão apenas pode ser realizada se existir um conhecimento sobre o estado dos dados existentes. Em suma, a qualidade dos dados em ambientes de DW deve ser assumida como um assunto merecedor da maior atenção nas organizações porque se trata duma condição necessária para o sucesso dos SDWs.

# Bibliografia


[Abreu, 1992] Abreu, F. "As Métricas na Gestão de Projectos de Desenvolvimento de Sistemas de Informação". 6as Jornadas para a Qualidade no Software, APQ. Lisboa. Dezembro, 1992.

[Adelman et al., 2005] Adelman, S., Moss, L. e Abai, M. "Data Strategy". Addison-Wesley Professional. 2005.

[Adelman, 2002] Adelman, S. "Measuring the Effectiveness of Your Data Warehouse". Principal Sid Adelman & Associates. 2002.

[Amaral & Varajão, 2000] Amaral, L. e Varajão, J. "Planeamento de Sistemas de Informação". FCA – Editora de Informática. 2000.

[Amaral et al., 2002] Amaral, L., Santos, L. e Oliveira, J. "Migração de Dados do Sistema Científico Português para a Plataforma Lattes". Departamento de Sistemas de Informação, Escola de Engenharia, Universidade do Minho. Guimarães, Portugal. 2002.

[Amaral, 2003] Amaral, G. "AQUAWARE: Um Ambiente de Suporte à Qualidade de Dados em Data Warehouse". Tese de Mestrado. Instituto de Matemática – Universidade Federal do Rio de Janeiro. Rio de Janeiro, Brasil. 2003.

[Ananthakrishna et al., 2002] Ananthakrishna, R., Chaudhuri, S. e Ganti, V. "Eliminating Fuzzy Duplicates in Data Warehouses". Proceedings of the 28th on Very Large Databases Conference. Hong Kong, China. 2002.

[Ballou & Tayi, 1998] Ballou, D. e Tayi, G. "Examining Data Quality". Communications of the ACM, vol. 41, nº2, pp. 54-57. Fevereiro, 1998.

[Ballou & Tayi, 1999] Ballou, D. e Tayi, G. "Enhancing Data Quality in Data Warehouse Environments". Communications of the ACM, vol. 42, nº1, pp. 73-78. Janeiro, 1999.

[Ballou et al, 2004] Ballou, D., Wang, R. e Madnick, S. "Special Section: Assuring Information Quality". Journal of Management Information Systems, vol. 20, nº3, pp. 9-11. 2004.

[Ballou et al., 1998] Ballou, D., Wang, R. Pazer, H. e Tayi, G. "Modeling Information Manufacturing Systems to Determine Information Product Quality". Management Science nº44, pp. 462-484. 1998.

[Basili et al., 1994] Basili, V., Caldiera, G. e Rombach, H. "The Goal Question Metric Approach". Enciclopédia de Engenharia de Software 2º volume, pp. 528-532. John Wiley & Sons. Inc. 1994.

[Bobrowski et al., 1998] Bobrowski, M., Marré, M., Yankelevich, D. "A Software Engineering View of Data Quality". Departamento de Computação, Faculdade de Ciências Exactas e Naturais, Universidade de Buenos Aires. Buenos Aires, Argentina. 1998.

[Bobrowski et al., 1999] Bobrowski, M., Marré, M., Yankelevich, D. "Measuring Data Quality". Relatório Técnico. Departamento de Computação, Faculdade de Ciências Exactas e Naturais, Universidade de Buenos Aires. Buenos Aires, Argentina. 1999.

[Bouzeghoub & Peralta, 2004] Bouzeghoub, M. e Peralta, V. "A Framework for Analysis of Data Freshness". Communications of the ACM, pp. 59-67. Maison de la Chimie, Paris, França. 2004.

[Bouzeghoub et al., 1999] Bouzeghoub, M., Fabret, F. e Broqué, M. "Modeling Data Warehouse Refreshment Process as a Workflow Application". Proceedings of the International Workshop on Design and Management of Data Warehouses (DMDW'99). 1999.

[Brackett, 1996] Brackett, M. "The Data Warehouse Challenge". New York: John Wiley & Sons, Inc. 1996.

[Brackstone, 1999] Brackstone, G. "Managing Data Quality in a Statistical Agency". Survey Methodology, vol. 25, nº2, pp.139-149. Canadá. Dezembro, 1999.

[Brackstone, 2001] Brackstone, G. "How important is accuracy?". Proceedings of Statistics Canada Symposium. Canadá. 2001.

[Calero et al., 2001] Calero, C., Piattini, M., Pascual, C. e Serrano, M. "Towards Data Warehouse Quality Metrics". Proceedings of the 3th International Workshop on Design and Management of Data Warehouses (DMDW'01). Interlaken, Suiça. Junho, 2001.

[Cantone & Donzelli, 1998] Cantone, G., e Donzelli P. "Developing and Maintaining Software Measurement Models". International Software Engineering Research Network. Technical Report, 98-33. 1998.

[Cantone & Donzelli, 1999] Cantone, G., e Donzelli P. "Production and Maintenance of Software Measurement Models". International Software Engineering Research Network. Technical Report, 99-19. 1999.

[Cappiello & Francalanci, 2002] Cappiello, C. e Francalanci, C. "Considerations about Costs Deriving from a Poor Data Quality". Relatório do Projecto DaQuinCIS. 2002.

[Cappiello et al., 2004] Cappiello, C., Francalanci, C. e Pernici, B. "Data quality assessment from the user's perspective". Communications of the ACM, pp. 59-67. Maison de la Chimie, Paris, França. 2004.

[Carvalho, 2003] Carvalho, F. "Mineração de dados – Problemática geral da preparação dos dados". Aulas de Mineração de Dados. www.di.ufpe.br. 2003.

[Chaudhuri & Dayal, 1997] Chaudhuri, S. e Dayal, U. "An Overview of Data Warehousing and OLAP Technology". SIGMOD Record, vol. 26, nº1. Março, 1997.

[Chung et al, 2002] Chung, W., Fisher, C. e Wang, R. “What Skills Matter in Data Quality?”. Information Conference on Information Quality. Massachusetts Institute of Technology. Novembro, 2002.

[Copeland & Simpson, 2004] Copeland, C. e Simpson, M. ”The Information Quality Act: OMB’s Guidance and Initial Implementation”. Congressional Research Service – The Library of Congress. Agosto, 2004.

[Cordeiro, 2004] Cordeiro, J. “Reflexões sobre a Gestão da Qualidade Total: fim de mais um modismo ou incorporação do conceito por meio de novas ferramentas de gestão?”. Revista da FAE, vol. 7, nº1, p.23-33. Curitiba, Brasil. Janeiro/Junho, 2004.

[Dataflux, 1999] Dataflux. “Data Quality”. White Paper. DataFlux Corporation, www.dataflux.com. 1999.

[Dataflux, 2004] DataFlux, “Data Monitor – Taking Control of Your Information Assets”. White Paper. DataFlux Corporation, www.dataflux.com. 2004.

[Eckerson, 2002] Eckerson, W. “Data Quality and The Bottom Line: Achieving Business Success Through a Commitment to High Quality Data”. The Data Warehousing Institute. www.dw-institute.com. Seattle, EUA. 2002.

[English, 1999] English, L. “Improving Data Warehouse and Business Information Quality”. John Wiley & Sons, Inc. New York, EUA. 1999.

[English, 2001] English, L. “Ten Years of Information Quality Advances: What Next?”. Information Impact International, Inc. 2001.

[English, 2002a] English, L. “Mistakes to avoid if your Data Warehouse is to deliver Quality Information”. Information Impact International, Inc. 2002.

[English, 2002b] English, L. “Ten Essentials of Information Quality Management. Information. Information” Impact International, Inc. 2002.

[English, 2003a] English, L. “How to Save $576 925 000 Through IQ Management”. Information Impact International, Inc. 2003.

[English, 2003b] English, L. “Total Informaton Quality Management: A Complete Methodology for IQ Management”. Information Impact International, Inc. 2003.

[English, 2003c] English, L. “The Information Quality Act: Mandate for IQ”. Information Impact International, Inc. 2003.

[English, 2004] English, L. “Data Quality - Standardize, Validate and Improve your Information Assets”. DataFlux Corporation www.dataflux.com. 2004.

[Fabret et al., 1997] Fabret, F., Matulovic, M. e Simon, E. “State of the Art: Data Warehouse Refreshment”. Foundations of Data Warehouse Quality (DWQ). DWQ Consortium. 1997.

[Fischer & Kingma, 2001] Fischer, C. e Kingma, B. “Critically of DQ as exemplified in two disasters”. Information & Management, vol. 39, pp. 109-116. 2001.

[Galhardas et al., 2000] Galhardas, H., Florescu, D., Shasha, D. e Simon, E. “AJAX: an Extensible Data Cleaning Tool”. Proceedings of the 2000 ACM SIGMOD International Conference on Management of data. Dallas, Texas, EUA. Maio, 2000.

[Galhardas et al., 2001] Galhardas, H., Florescu, D., Shasha, D. e Simon, E. “Improving Data Cleaning Quality using a Data Lineage Facility”. Proceedings of the 3th International Workshop on Design and Management of Data Warehouses (DMDW’2001). Interlaken, Suiça. Junho, 2001.

[Ganhão, 1994] Ganhão, F. “Gestão da Qualidade – Área da Produção”. IAPMEI. 1994.

[Gonzales, 2003] Gonzales, M. “Enterprise Data Quality for Business Intelligence”. The Focus Group, Ltd. Outubro, 2003.

[Gonzales, 2004] Gonzales, M. “The Architecture of Enterprise Data Quality”. www.intelligententerprise.com. Junho, 2004.

[Graville, 2004] Graville, D. “Measurement and Metrics”. Quarterly Newsletter of Fortna, vol. 6, issue 4. Setembro/Dezembro, 2004.

[Helfert & Herrmann, 2002] Helfert, M. e Herrmann, C. “Proactive Data Quality Management for Data Warehouse Systems - A Metadata based Data Quality System”. Competence Center .Data Warehousing 2 (CCDW2). Institute of Information Management, University of St. Gallen. Suiça. 2002.

[Helfert & Maur, 2001] Helfert, M. & Maur, E. “A Strategy for Managing Data Quality in Data Warehouse Systems”. Institute of Information Management, University of St. Gallen. Suiça. 2001.

[Helfert & Radon, 2000] Helfert, M. e Radon. “An Approach for Information Quality measurement in Data Warehousing”. Proceedings of the 2000 Conference on Information Quality, pp. 109-125. Massachusetts Institute of Technology. Cambridge, EUA. 2000.

[Helfert, 2001] Helfert, M. “Managing and Measuring Data Quality in Data Warehousing”. Proceedings of the World Multiconference on Systemics, Cybernetics and Informatics, pp. 55-65. Orlando, EUA. 2001.

[Helfert, 2003] Helfert, M. “Data Quality Management for Data Warehouse Systems”. School of Computing, Dublin City University. Dublin, Irlanda. 2003.

[Hernández & Stolfo, 1995] Hernández, M. e Stolfo, S. “The Merge/Purge Problem for Large Databases”. Proceedings of the ACM SIGMOD Conference. 1995.

[Hernandez & Stolfo, 1998] Hernandez, M. e Stolfo, S. “Real-World Data is Dirty: Data Cleansing and the Merge/Purge Problem”. Journal of Data Mining and Knowledge Discovery, 2(1), pp. 9-37. Department of Computer Science, Columbia University. New York, EUA. 1998.

[Hudicka, 2002] Hudicka, J. “Develop a Data Quality Strategy Before Implementing a Data Warehouse”. DM Review Magazine. 2002.

[Hussain & Beg, 2003] Hussain, S. e Beg, J. “Data Quality – A problem and An Approach”. Wipro Technologies. 2003.

[Igor & Mahnic, 2000] Igor, R. e Mahnic, V. “Data Quality: A Prerequisite for Successful Data Warehouse Implementation”. IT and Networking Security. 2000.

[Inmon et al., 1998] Inmon, W., Rudin, K., Buss, C. e Sousa, R. “Data Warehouse Performance”. John Wiley & Sons, Inc. New York, EUA. 1998.

[Inmon, 1996] Inmon, W. “Building the Data Warehouse”. Second Edition. John Wiley & Sons, Inc. New York, EUA. 1996.

[Inmon, 2006a] Inmon, W. “Metadata Management”. Business Intelligence Network. Março, 2006.

[Inmon, 2006b] Inmon, B. “An Introduction to DWs 2.0”. Business Intelligence Network. Março, 2006.

[Iwaysoftware, 2004] Iwaysoftware. “iway – Data Management Solutions”. www.iwaysoftware.com. 2004.

[Jarke & Vassiliou, 1997] Jarke, M. e Vassiliou, Y. “Data Warehouse Quality: A Review of the DWQ Project”. Proceedings 2nd Conference on Information Quality. Massachusetts Institute of Technology. Cambridge, EUA. 1997.

[Jarke et al., 1999] Jarke, M., Jeusfeld, M., Quix, C. e Vassiliadis, P. “Architecture and Quality in Data Warehouses: An Extended Repository Approach”. Information Systems, vol. 24, nº3, pp. 229-253. 1999.

[Jarke et al., 2003] Jarke, M., Grimmer, U. e Dominik, L. “Systematic Development of Data Mining-Based Data Quality Tools”. Proceedings of the 29th Very Large Databases Conference. Berlim, Alemanha. 2003.

[Jeusfeld et al., 1998] Jeusfeld, M., Quix, C. e Jarke, M. “Design and Analysis of Quality Information for Data Warehouses”. Proceedings of the 17th International Conference on the Entity Relationship Approach (ER'98). Singapore. 1998.

[Kahn et al., 2002] Kahn, B., Strong, D. e Wang, R. ”Information Quality Benchmarks: Product and Service Performance”. Communications of the ACM, vol. 45, nº4, pp. 184-192. Abril, 2002.

[Kenyon et al., 2000] Kenyon, H., Cassella, J., Lambert, R. e Jordaan, W. “Global Data Management Survey - the new economy is the data economy”. PriceWaterhouseCoopers. www.pwc.com. 2000.

[Kenyon et al., 2004] Kenyon, H., Smith, P., Jordaan, W., Marinos, G. e Benge, J. “Data Quality Management”. PriceWaterhouseCoopers. www.pwc.com. 2004.

[Kim et al., 2003] Kim, W., Choi, B., Hong, E., Kim, S. e Lee, D. “A Taxonomy of Dirty Data”. Data Mining and Knowledge Discovery, nº7, pp. 81-99. Kluwer Publishers. 2003.

[Kimball & Caserta, 2004] Kimball, R. e Caserta, J. “The Data Warehouse ETL Toolkit: Practical Techniques for Extracting and Cleaning”. John Wiley & Sons, Inc. New York, EUA. 2004.

[Kimball et al., 1998] Kimball, R., Reeves, L., Ross, M. e Thornthwaite, W. “The Data Warehouse Lifecycle Toolkit”. John Wiley & Sons, Inc. New York, EUA. 1998.

[Kimball, 2004] Kimball, R. “Surprising Value of Data Profiling”. Kimball Design Tip nº 59. Kimball University. Setembro, 2004.

[Kyl, 2005] Kyl, J. “The Data Quality Act: History and Purpose”. United States Senade - Republican Policy Committee. Janeiro, 2005.

[Lee & Strong, 2003] Lee. Y. e Strong, D. “Process Knowledge and Data Quality Outcomes”. Proceedings of the 8th International Conference on Information Quality (ICIQ-03). 2003.

[Lee & Strong, 2004] Lee, Y. e Strong, D. “Knowing-Why About Data Processes and Data Quality”. Journal of Management Information Systems, vol. 20, nº 3, pp. 13-39. 2004.

[Lee et al., 2000a] Lee, M., Ling, T. e Low, W. “IntelliClean: A Knowledge-Based Intelligent Data Cleaner”. Proceedings of the ACM SIGKDD. Boston, EUA. 2000.

[Lee et al., 2000b] Lee, Y., Jarke, M., Madnick, S., Wand, Y., Funk, J. e Bowen, P. “Data Quality in Internet Time, Space, and Communities”. pp. 713-716. 2000.

[Lee et al., 2002] Lee, Y., Strong, D., Kahn, B., e Wang, R. “AIMQ: A Methodology for Information Quality Assessment”. Elsevier Science. Information & Management, vol.40, nº2, pp.133-146. 2002.

[Leitheiser, 2001] Leitheiser, R. “Data Quality in Health Care Data Warehouse Environments”. Proceedings of the 34th International Conference on System Sciences. IEEE. 2001.

[Low et al., 2001] Low, W., Lee, M. e Ling, T. “A knowledge-based approach for duplicate elimination in data cleaning”. Elsevier Science. Information Systems, nº 26, pp. 505-606. 2001.

[Loshin, 2005] Loshin, D. “Developing Information Quality Metrics”. DM Review Magazine. Maio, 2005.

[Marco, 2004] Marco, D. “Designing the Optimal Meta Data Tool - Part One”. Enterprise Warehousing Solutions, Inc. www.tdan.com. Outubro, 2004.

[Marcus & Maletic, 2000] Marcus, A. e Maletic, J. “Data Cleansing: Beyond Integrity Analysis”. Proceedings of the Information Quality. Junho, 2000.

[Marques, 1994] Marques, H. “Sebenta de Gestão da Produção”. Instituto Superior de Línguas e Administração. Outubro, 1994.

[McKnight, 2003] McKnight, W. “Overall Approach to Data Quality ROI”. White Paper. Firstlogic, Inc. 2003.

[Mendes, 2006] Mendes. A. “Entrevistas no âmbito projecto Metricware, sobre o problema da qualidade dos dados no DM de vendas do caso de estudo”. Entrevistador: Alexandre Costa. Janeiro/Março, 2006.

[Moreira, 2001] Moreira, A. “Gestão pela Qualidade Total”. Universidade de Aveiro. 2001.

[Müller & Freytag, 2002] Müller, H., Freytag, J. “Problems, Methods and Challenges in Comprehensive Data Cleansing”. Humboldt – Universität zu Berlin zu Berlin. Berlim, Alemanha. 2002.

[Naumann & Rolker, 2000] Naumann, F. e Rolker, C. “Assessment Methods for Information Quality Criteria”. German Research Society. 2000.

[Naumann & Roth, 2004] Naumann, F. e Roth, M. “Information Quality: How Good are Off-The-Shelf DBMS?”. 2004.

[Neely, 1998] Neely, M. “Data Quality Tools for Data Warehousing – A Small Sample Survey”. Center for Technology, University of Albany. EUA. 1998.

[Nguyen & Fisher, 2000] Nguyen, T. e Fisher, T. “The case for ETL – Merging Data Warehousing and Data Quality, because bad data costs companies money”. White paper. DM Review Magazine. 2000.

[Novabase, 2002] Novabase. “Qualidade dos dados On-Line”. White Paper. Novabase. 2002.

[Oliveira et al., 2004] Oliveira, P., Henriques, P. e Rodrigues, F. “Limpeza de Dados – Uma Visão Geral”. Proceedings of Data Gadgets' 2004 Workshop – Bringing Up Emerging Solutions for Data Warehousing Systems, pp. 39-51. Málaga, Espanha. Novembro, 2004.

[Oliveira et al., 2005a] Oliveira, P., Rodrigues, F., Henriques, P. e Galhardas, H. “A Taxonomy of Data Quality Problems”. Proceedings of 2nd International Workshop on Data and Information Quality. (in conjuction with CAISE'05), pp. 219-233. Porto, Portugal. Junho, 2005.

[Oliveira et al., 2005b] Oliveira, P., Rodrigues, F. e Henriques, P. “A Framework for Detection and Correction of Data Quality Problems”. Proceedings of Data Gadgets' 2005 Workshop – Bringing Up Emerging Solutions for Data Warehousing Systems. pp. 50-74. Granada, Espanha. Setembro, 2005.

[Olson, 2003] Olson, J. “Data Quality: The Accuracy Dimension”. Morgan Kaufmann Publishers. Elsevier Science. 2003.

[Orr, 1998] Orr, K. (1998). Data Quality and Systems Theory. Communications of the ACM, vol.41, nº2, pp. 66-71. Fevereiro, 1998.

[Parssian et al., 1999] Parssian, A., Sarkar, S. e Jacob, V. “Assessing Data Quality for Information Products”. School of Management, University of Texas. Dallas, EUA. 1999.

[Paulson, 2000] Paulson, L. “Data Quality: A Rising E-Business Concern”. IT Pro. Julho/Agosto, 2000.

[Piattini et al., 2001] Piattini, M., Genero, M. e Calero, C. “Data Model Metrics”. Handbook of Software Engineering and Knowledge Engineering. 2001.

[Pierce, 2004a] Pierce, E. “Developing, Implementing and Monitoring an Information Product Quality Strategy”. Proceedings of 9th International Conference on Information Quality (ICIQ-04). 2004.

[Pierce, 2004b] Pierce, E. “Assessing DQ with Control Matrices”. Communications of the ACM, vol.47, nº2, pp. 82-86. Fevereiro, 2004.

[Pipino et al., 2002] Pipino, L., Lee, Y. e Wang. R. "Data Quality Assessment". Communications of the ACM, vol.45, nº4, pp. 211-218. Abril, 2002.

[PMBoK, 2000] Project Management Institute, "PMBoK Guide - A Guide to the Project Management Body of Knowledge". 2000.

[Rahm & Do, 2000] Rahm, E. e Do, H. "Data Cleaning: Problems and Current Approaches". IEEE Bulletin of the Technical Committee on Data Engineering, 24, 4. 2000.

[Raisinghani, 1999] Raisinghani, V. "Cleaning Methods in Data Warehousing". School of Information Technology. Bombay. 1999.

[Raman & Hellerstein, 2001] Raman, V., Hellerstein, J. "Potter's Wheel: An Interactive Data Cleansing System". Proceedings of the 27th on Very Large Databases Conference. Rome, Italy. 2001.

[Rascão, 2000] Rascão, J. "A Análise Estratégica e o Sistema de Informação para a Tomada de Decisão Estratégica". Editora Sílabo. Lisboa, Portugal. 2000.

[Rascão, 2001] Rascão, J. "Sistemas de Informação para as Organizações". Editora Sílabo. Lisboa, Portugal. 2001.

[Rasmussen, 2004] Rasmussen, K. "Elementary Data Quality Elements". IASSIST 2004. Madison, USA. 2004.

[Redman, 1995] Redman, T. "Improve Data Quality for Competitive Advantage". Sloan Management Review. Cambridge. Winter, 36, 2, 99. 1995.

[Redman, 1996] Redman, T. "Data quality for the information age". Norwood, Artech House. 1996.

[Redman, 1998] Redman, T. "The Impact of Poor Data Quality on the Typical Enterprise". Communications of the ACM, vol.41, nº2, pp.79-82. Fevereiro, 1998.

[Redman, 2004] Redman, T. "Data: An Unfolding Quality Disaster". DM Review Magazine. Agosto, 2004.

[Scannapieco & Catarci, 2002] Scannapieco, M e Catarci, T. "Data Quality under the Computer Science Perspective". Dipartimento di Informatica e Sistemistica, Universita di Rome. Roma, Itália. 2002.

[Scannapieco et al., 2003] Scannapieco M., Mecella M., Catarci T., Cappiello C., Pernici B., Mazzoleni F., Stella F. "Comparative Analysis of the Proposed Methodologies for Measuring and Improving Data Quality and Description of an Integrated Proposal". Relatório do Projecto DaQuinCIS. Dipartimento di Informatica e Sistemistica, Universita di Rome. Roma, Itália. 2002.

[Serrano & Filho, 2003] Serrano, A. e Filho, C. Gestão do Conhecimento. FCA – Editora de Informática. 2003.

[Serrano et al., 2002] Serrano, M., Calero, C. e Piattini, M. "Experimental Validation of Multidimensional Data Models Metrics". Proceedings of the 36th Hawaii International Conference on Systems Sciences (HICSS'03). IEEE. Computer Society. 2002.

[Serrano et al., 2003] Serrano, M., Calero, C., Piattini, M. e Caballero, I. "Calidad de los Almacenes de Datos". I+D Computación, vol. 2, nº2. Julho, 2003.

[Shankaranarayan et al., 2000] Shankaranarayan, G., Ziad, M. e Wang, R. "IP-MAP: Representing the Manufacture of an Information Product". Proceedings of the International Conference on Information Quality (ICIQ-00). 2000.

[Shankaranarayan et al., 2003] Shankaranarayan, G., Ziad, M., Wang. R. "Managing Data Quality in Dynamic Decision Environments: An Information Product Approach". Journal of Database Management, vol. 14, nº4. Outubro/Dezembro, 2003.

[Shankaranarayan, 2005] Shankaranarayan, G. "Towards Implementing Total Data Quality Management in a Data Warehouse". Journal of Information Technology Management, vol. 16, nº 1. 2005.

[Silva, 2003] Silva, M. "Dicionário Terminológico da Gestão pela Qualidade Total em Serviços". Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo. Ph.D. Thesis. 2003.

[Skriletz, 2002] Skriletz, R. "Strategic Insight: Data Quality: Fact and Perception". DM Review Magazine. Novembro, 2002.

[Smith, 2004a] Smith, G. "A Primer on Metrics". www.intelligententerprise.com. Março, 2004.

[Smith, 2004b] Smith, G. "A Primer on Metrics, Part Two". www.intelligententerprise.com. Março, 2004.

[Strong et al., 1997] Strong, D., Lee, Y. e Wang, R. "Data Quality in Context". Communications of the ACM, vol.40, nº5, pp. 103-110. Maio, 1997.

[Theodoratos & Bouzeghoub, 1999] Theodoratos, D. e Bouzeghoub, M. "Data Currency Quality Factors in Data Warehouse Design". Proceedings of the International Workshop on Design and Management of Data Warehouses (DMDW'99). Heidelberg. Alemanha. Junho, 1999.

[Tzu, 1994] Tzu, S. "A Arte da Guerra". Publicações Europa-América, Lda. 1994.

[Varajão, 1998] Varajão, J. "Arquitectura da Gestão de Sistemas de Informação". FCA – Editora de Informática. 1998.

[Vassiliadis et al., 1999] Vassiliadis, P., Bouzeghoub, M. e Quix, C. "Towards Quality-Oriented Data Warehouse Usage and Evolution". Proceedings 11th Conference of Advanced Information Systems Engineering (CAiSE '99), Heidelberg, Germany. 1999.

[Vassiliadis et al., 2001] Vassiliadis, P., Vagena, Z., Skiadopoulos, S., Karayannidis, N. e Sellis, T. "ARKTOS: A Tool for Data Cleaning and Transformation in Data Warehouse Environments". Information Systems, vol. 26, pp. 537-561. 2001.

[Vassiliadis, 2000] Vassiliadis, P. "Data Warehouse Modeling and Quality Issues". Department of Electrical and Computer Engineering, National Technical University of Athens. Ph.D. Thesis. Zographou, Grécia. 2000.

[Wand & Wang, 1996] Wand, Y., Wang, R. “Anchoring Data Quality Dimensions in Ontological Foundations”. Communications of the ACM, vol.39, nº11, pp. 86-95. Novembro, 1996.

[Wang et al, 1994] Wang, R., Strong, D. e Guarascio, L. “Beyond Accuracy: What Data Quality Means to Data Consumers”. TDQM Research Program, Sloan School of Management, Massachusetts Institute of Technology. Cambridge, EUA. Outubro, 1994.

[Wang et al., 1998] Wang, R., Lee, Y., Pipino, L. e Strong, D. “Manage Your Information as a Product”. Sloan Management Review, vol. 39, pp. 95-105. 1998.

[Wang et al., 2003] Wang, R., Madnick, S., Harris, W. e Allen, T. “An Information Product Approach for Total Information Awareness”. IEEE Aerospace Conference. Montana, EUA. 2003.

[Wang et al., 2004] Wang, R., Lee, Y. e Davidson, B. “Developing data production maps: meeting patient discharge data submission requirements”. Healthcare Technology and Management, vol. 6, nº2. Inderscience Enterprises Ltd. Julho, 2004.

[Wang, 1998] Wang, R. “A Product Perspective on Total Quality Management”. Communications of the ACM, vol.41, nº2, pp.58-65. Fevereiro, 1998.

[Wang, 2004] Wang, R. “Data Quality: Theory in Practice”. EPA 23rd Annual National Conference on Managing Environmental Quality Systems. 2004.

[Watson et al., 2001] Watson, H., Wixom, B., Annino, D., Avery, K. e Rutherford, M. “Current Practices in Data Warehousing”. Data Warehouse Today. Information Systems Management. 2001.

[Watson et al., 2002] Watson, H., Goodhue, D. e Wixom, B. “The Benefits of Data Warehouse: Why some organizations realize exceptional payoffs”. Information & Management, vol. 39, pp. 491-502. Elsevier Science. 2002.

[White, 2000] White, C. “An Analysis-Led Approach to Data Warehouse Design and Development. Database Associates”. White Paper. Database Associates. Janeiro, 2000.

[Wood, 2002] Wood, G. “Standardization of data quality metrics”. Computing & Control Engineering Journal. Outubro, 2002.

# Referências WWW

[1] www.egi.ua.pt

Este sítio fornece uma elevada quantidade de informação acerca do conceito de qualidade, bem como, os instrumentos e as metodologias para a sua gestão. Acedido em 16 de Fevereiro de 2005.

[2] http://rpc.senate.gov

Este sítio reporta ao senado norte-americano e fornece informações relativas às razões e objectivos do DQA. Acedido em 1 de Março de 2005.

[3] www.uschamber.com

O sítio da Casa de Comércio dos E.U.A. disponibiliza um conjunto de documentos sobre problemas derivados da fraca qualidade dos dados e refere medidas governamentais e sectoriais que visam elevar os critérios de qualidade dos dados disseminados pelas organizações. Acedido em 28 de Fevereiro de 2005.

[4] http://www.whitehouse.gov/omb/egov/

Este sítio é referente ao OMB, sob alçada do gabinete executivo do Presidente dos E.U.A. e disponibiliza, no âmbito do OMB, informações relativas à divulgação dos dados entre as agências federais, nomeadamente, as orientações chave a respeitar de modo a assegurar e maximizar a qualidade, objectividade, utilidade e integração da informação divulgada. Acedido em 1 de Março de 2005.

[5] http://www.smith-robertson.com

Este sítio disponibiliza informações relativas às orientações sobre a iniciativa DQA. Acedido em 28 de Fevereiro de 2005.

[6] http://www.b-eye-network.com

Este sítio disponibiliza informações e conferências *on-line* relacionadas com as últimas tendências dos assuntos que cobrem diversos temas em SDWs. Acedido em 20 de Abril de 2006.

[7] http://web.tagus.ist.utl.pt/~helena.galhardas/cleaning.html

Este sítio disponibiliza informações relativas a um conjunto alargado de *software* para o tratamento da qualidade dos dados. Acedido em 20 de Outubro de 2005.

[8] http://www.dbstar.com

Sítio da empresa *Evoke Software*, que presta informações relativamente às aplicações de tratamento dos dados *AXIO* e *ATHANOR*. Acedido em 10 de Novembro de 2005.

[9] http://www.wizsoft.com/index.html

Sítio da empresa *WizSoft Inc.*, que disponibiliza informações relativas ao *software WizWhy*, *WizRule* e *Wizsame*, usados no estudo de caso. Acedido em 10 de Novembro de 2005.

[10] http://www.trilliumsoft.com/trilliumsoft.nsf

Sítio da empresa Trillium Software, que fornece informações sobre as aplicações Trillium Software System e Trillium Software Discovery. Acedido em 10 de Novembro de 2005.

[11] http://www.msi.com.au/

Sítio da empresa *Group1 Software*, que fornece informações sobre a aplicação *NADIS*. Acedido em 10 de Novembro de 2005.

[12] http://www.helpit.co.uk/

Sítio da empresa *helpIT Systems Limited*, disponibilizando informações sobre a aplicação *matchIT*. Acedido em 10 de Novembro de 2005.

[13] http://www.g1.com/

Sítio da empresa *Group1 Software*, que fornece informações sobre a aplicação *Merge/Purge Plus*. Acedido em 10 de Novembro de 2005.

[14] www.dataflux.com

Sítio da empresa *DataFlux Corporation*, que fornece informações sobre a aplicação *dfPower Studio*. Acedido em 10 de Novembro de 2005.

[15] www.teradata.com

Sítio da empresa Teradata (NCR) , que presta informações sobre as aplicações Teradata Warehouse e Teradata Warehouse Miner. Acedido em 10 de Novembro de 2005.

[16] www.datiris.com.

Sítio da empresa *Datiris*, fornecendo informações sobre a aplicação de *data profiling* usada no estudo de caso. Acedido em 26 de Dezembro de 2006.

[17] http://mitiq.mit.edu

Sítio do *Massachusetts Institute of Technology* que disponibiliza artigos sobre a qualidade dos dados, centrados essencialmente na TDQM. Acedido em 20 de Abril de 2004.

[18] http://www.inmoncif.com/home/

Este sítio disponibiliza um conjunto de artigos sobre os desafios, tendências e realidades dos SDWs, na óptica de um dos maiores impulsionadores destes sistemas, *Bill Inmon*. Acedido em 20 de Abril de 2004.